UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1996

SUMMER FACULTY RESEARCH PROGRAM FINAL REPORTS

VOLUME 4

ROME LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES

5800 Uplander Way

Culver City, CA 90230-6608

Program Director, RDL
Gary Moore

Program Manager, AFOSR
Major Linda Steel-Goodwin

Program Manager, RDL
Scott Licoscos

Program Administrator, RDL
Johnetta Thompson

Program Administrator, RDL
Rebecca Kelly

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Bolling Air Force Base

Washington, D.C.

December 1996

REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-00-
0731

d reviewing
Information

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REP(
	December, 1996	

4. TITLE AND SUBTITLE
1996 Summer Research Program (SRP), Summer Faculty Research Program (SFRP), Final Reports, Volume 4, Rome Laboratory

5. FUNDING NUMBERS
F49620-93-C-0063

6. AUTHOR(S)
Gary Moore

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Research & Development Laboratories (RDL)
5800 Uplander Way
Culver City, CA 90230-6608

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research (AFOSR)
801 N. Randolph St.
Arlington, VA 22203-1977

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The United States Air Force Summer Research Program (USAF-SRP) is designed to introduce university, college, and technical institute faculty members, graduate students, and high school students to Air Force research. This is accomplished by the faculty members (Summer Faculty Research Program, (SFRP)), graduate students (Graduate Student Research Program (GSRP)), and high school students (High School Apprenticeship Program (HSAP)) being selected on a nationally advertised competitive basis during the summer intersession period to perform research at Air Force Research Laboratory (AFRL) Technical Directorates, Air Force Air Logistics Centers (ALC), and other AF Laboratories. This volume consists of a program overview, program management statistics, and the final technical reports from the SFRP participants at the Rome Laboratory.

14. SUBJECT TERMS
Air Force Research, Air Force, Engineering, Laboratories, Reports, Summer, Universities, Faculty, Graduate Student, High School Student

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements.***

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C - Contract	**PR** - Project
G - Grant	**TA** - Task
PE - Program Element	**WU** - Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with....; Trans. of....; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD -	See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE -	See authorities.
NASA -	See Handbook NHB 2200.2.
NTIS -	Leave blank.

Block 12b. Distribution Code.

DOD -	Leave blank.
DOE -	Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
	Leave blank.
NASA -	Leave blank.
NTIS -	

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self- explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

PREFACE

Reports in this volume are numbered consecutively beginning with number 1. Each report is paginated with the report number followed by consecutive page numbers, e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3.

This document is one of a set of 16 volumes describing the 1996 AFOSR Summer Research Program. The following volumes comprise the set:

VOLUME	TITLE
1	Program Management Report
	Summer Faculty Research Program (SFRP) Reports
2A & 2B	Armstrong Laboratory
3A & 3B	Phillips Laboratory
4	Rome Laboratory
5A, 5B & 5C	Wright Laboratory
6	Arnold Engineering Development Center, Wilford Hall Medical Center and Air Logistics Centers
	Graduate Student Research Program (GSRP) Reports
7A & 7B	Armstrong Laboratory
8	Phillips Laboratory
9	Rome Laboratory
10A & 10B	Wright Laboratory
11	Arnold Engineering Development Center, United States Air Force Academy, Wilford Hall Medical Center, and Wright Patterson Medical Center
	High School Apprenticeship Program (HSAP) Reports
12A & 12B	Armstrong Laboratory
13	Phillips Laboratory
14	Rome Laboratory
15A&15B	Wright Laboratory
16	Arnold Engineering Development Center

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

SRP Final Report Table of Contents

INTRODUCTION

The Summer Research Program (SRP), sponsored by the Air Force Office of Scientific Research (AFOSR), offers paid opportunities for university faculty, graduate students, and high school students to conduct research in U.S. Air Force research laboratories nationwide during the summer.

Introduced by AFOSR in 1978, this innovative program is based on the concept of teaming academic researchers with Air Force scientists in the same disciplines using laboratory facilities and equipment not often available at associates' institutions.

The Summer Faculty Research Program (SFRP) is open annually to approximately 150 faculty members with at least two years of teaching and/or research experience in accredited U.S. colleges, universities, or technical institutions. SFRP associates must be either U.S. citizens or permanent residents.

The Graduate Student Research Program (GSRP) is open annually to approximately 100 graduate students holding a bachelor's or a master's degree; GSRP associates must be U.S. citizens enrolled full time at an accredited institution.

The High School Apprentice Program (HSAP) annually selects about 125 high school students located within a twenty mile commuting distance of participating Air Force laboratories.

AFOSR also offers its research associates an opportunity, under the Summer Research Extension Program (SREP), to continue their AFOSR-sponsored research at their home institutions through the award of research grants. In 1994 the maximum amount of each grant was increased from $20,000 to $25,000, and the number of AFOSR-sponsored grants decreased from 75 to 60. A separate annual report is compiled on the SREP.

The numbers of projected summer research participants in each of the three categories and SREP "grants" are usually increased through direct sponsorship by participating laboratories.

AFOSR's SRP has well served its objectives of building critical links between Air Force research laboratories and the academic community, opening avenues of communications and forging new research relationships between Air Force and academic technical experts in areas of national interest, and strengthening the nation's efforts to sustain careers in science and engineering. The success of the SRP can be gauged from its growth from inception (see Table 1) and from the favorable responses the 1996 participants expressed in end-of-tour SRP evaluations (Appendix B).

AFOSR contracts for administration of the SRP by civilian contractors. The contract was first awarded to Research & Development Laboratories (RDL) in September 1990. After

completion of the 1990 contract, RDL (in 1993) won the recompetition for the basic year and four 1-year options.

2. PARTICIPATION IN THE SUMMER RESEARCH PROGRAM

The SRP began with faculty associates in 1979; graduate students were added in 1982 and high school students in 1986. The following table shows the number of associates in the program each year.

YEAR	SRP Participation, by Year			TOTAL
	SFRP	GSRP	HSAP	
1979	70			70
1980	87			87
1981	87			87
1982	91	17		108
1983	101	53		154
1984	152	84		236
1985	154	92		246
1986	158	100	42	300
1987	159	101	73	333
1988	153	107	101	361
1989	168	102	103	373
1990	165	121	132	418
1991	170	142	132	444
1992	185	121	159	464
1993	187	117	136	440
1994	192	117	133	442
1995	190	115	137	442
1996	188	109	138	435

Beginning in 1993, due to budget cuts, some of the laboratories weren't able to afford to fund as many associates as in previous years. Since then, the number of funded positions has remained fairly constant at a slightly lower level.

3. RECRUITING AND SELECTION

The SRP is conducted on a nationally advertised and competitive-selection basis. The advertising for faculty and graduate students consisted primarily of the mailing of 8,000 52-page SRP brochures to chairpersons of departments relevant to AFOSR research and to administrators of grants in accredited universities, colleges, and technical institutions. Historically Black Colleges and Universities (HBCUs) and Minority Institutions (MIs) were included. Brochures also went to all participating USAF laboratories, the previous year's participants, and numerous individual requesters (over 1000 annually).

RDL placed advertisements in the following publications: *Black Issues in Higher Education, Winds of Change,* and *IEEE Spectrum.* Because no participants list either *Physics Today* or *Chemical & Engineering News* as being their source of learning about the program for the past several years, advertisements in these magazines were dropped, and the funds were used to cover increases in brochure printing costs.

High school applicants can participate only in laboratories located no more than 20 miles from their residence. Tailored brochures on the HSAP were sent to the head counselors of 180 high schools in the vicinity of participating laboratories, with instructions for publicizing the program in their schools. High school students selected to serve at Wright Laboratory's Armament Directorate (Eglin Air Force Base, Florida) serve eleven weeks as opposed to the eight weeks normally worked by high school students at all other participating laboratories.

Each SFRP or GSRP applicant is given a first, second, and third choice of laboratory. High school students who have more than one laboratory or directorate near their homes are also given first, second, and third choices.

Laboratories make their selections and prioritize their nominees. AFOSR then determines the number to be funded at each laboratory and approves laboratories' selections.

Subsequently, laboratories use their own funds to sponsor additional candidates. Some selectees do not accept the appointment, so alternate candidates are chosen. This multi-step selection procedure results in some candidates being notified of their acceptance after scheduled deadlines. The total applicants and participants for 1996 are shown in this table.

1996 Applicants and Participants			
PARTICIPANT CATEGORY	TOTAL APPLICANTS	SELECTEES	DECLINING SELECTEES
SFRP	572	188	39
(HBCU/MI)	(119)	(27)	(5)
GSRP	235	109	7
(HBCU/MI)	(18)	(7)	(1)
HSAP	474	138	8
TOTAL	1281	435	54

4. SITE VISITS

During June and July of 1996, representatives of both AFOSR/NI and RDL visited each participating laboratory to provide briefings, answer questions, and resolve problems for both laboratory personnel and participants. The objective was to ensure that the SRP would be as constructive as possible for all participants. Both SRP participants and RDL representatives found these visits beneficial. At many of the laboratories, this was the only opportunity for all participants to meet at one time to share their experiences and exchange ideas.

5. HISTORICALLY BLACK COLLEGES AND UNIVERSITIES AND MINORITY INSTITUTIONS (HBCU/MIs)

Before 1993, an RDL program representative visited from seven to ten different HBCU/Mis annually to promote interest in the SRP among the faculty and graduate students. These efforts were marginally effective, yielding a doubling of HBCI/MI applicants. In an effort to achieve AFOSR's goal of 10% of all applicants and selectees being HBCU/MI qualified, the RDL team decided to try other avenues of approach to increase the number of qualified applicants. Through the combined efforts of the AFOSR Program Office at Bolling AFB and RDL, two very active minority groups were found, HACU (Hispanic American Colleges and Universities) and AISES (American Indian Science and Engineering Society). RDL is in communication with representatives of each of these organizations on a monthly basis to keep up with the their activities and special events. Both organizations have widely-distributed magazines/quarterlies in which RDL placed ads.

Since 1994 the number of both SFRP and GSRP HBCU/MI applicants and participants has increased ten-fold, from about two dozen SFRP applicants and a half dozen selectees to over 100 applicants and two dozen selectees, and a half-dozen GSRP applicants and two or three selectees to 18 applicants and 7 or 8 selectees. Since 1993, the SFRP had a two-fold applicant

increase and a two-fold selectee increase. Since 1993, the GSRP had a three-fold applicant increase and a three to four-fold increase in selectees.

In addition to RDL's special recruiting efforts, AFOSR attempts each year to obtain additional funding or use leftover funding from cancellations the past year to fund HBCU/MI associates. This year, 5 HBCU/MI SFRPs declined after they were selected (and there was no one qualified to replace them with). The following table records HBCU/MI participation in this program.

SRP HBCU/MI Participation, By Year				
YEAR	SFRP		GSRP	
	Applicants	Participants	Applicants	Participants
1985	76	23	15	11
1986	70	18	20	10
1987	82	32	32	10
1988	53	17	23	14
1989	39	15	13	4
1990	43	14	17	3
1991	42	13	8	5
1992	70	13	9	5
1993	60	13	6	2
1994	90	16	11	6
1995	90	21	20	8
1996	119	27	18	7

6. SRP FUNDING SOURCES

Funding sources for the 1996 SRP were the AFOSR-provided slots for the basic contract and laboratory funds. Funding sources by category for the 1996 SRP selected participants are shown here.

1996 SRP FUNDING CATEGORY	SFRP	GSRP	HSAP
AFOSR Basic Allocation Funds	141	85	123
USAF Laboratory Funds	37	19	15
HBCU/MI By AFOSR (Using Procured Addn'l Funds)	10	5	0
TOTAL	188	109	138

SFRP - 150 were selected, but nine canceled too late to be replaced.
GSRP - 90 were selected, but five canceled too late to be replaced (10 allocations for the ALCs were withheld by AFOSR.)
HSAP - 125 were selected, but two canceled too late to be replaced.

7. COMPENSATION FOR PARTICIPANTS

Compensation for SRP participants, per five-day work week, is shown in this table.

1996 SRP Associate Compensation

PARTICIPANT CATEGORY	1991	1992	1993	1994	1995	1996
Faculty Members	$690	$718	$740	$740	$740	$770
Graduate Student (Master's Degree)	$425	$442	$455	$455	$455	$470
Graduate Student (Bachelor's Degree)	$365	$380	$391	$391	$391	$400
High School Student (First Year)	$200	$200	$200	$200	$200	$200
High School Student (Subsequent Years)	$240	$240	$240	$240	$240	$240

The program also offered associates whose homes were more than 50 miles from the laboratory an expense allowance (seven days per week) of $50/day for faculty and $40/day for graduate students. Transportation to the laboratory at the beginning of their tour and back to their home destinations at the end was also reimbursed for these participants. Of the combined SFRP and

GSRP associates, 65 % (194 out of 297) claimed travel reimbursements at an average round-trip cost of $780.

Faculty members were encouraged to visit their laboratories before their summer tour began. All costs of these orientation visits were reimbursed. Forty-five percent (85 out of 188) of faculty associates took orientation trips at an average cost of $444. By contrast, in 1993, 58 % of SFRP associates took orientation visits at an average cost of $685; that was the highest percentage of associates opting to take an orientation trip since RDL has administered the SRP, and the highest average cost of an orientation trip. These 1993 numbers are included to show the fluctuation which can occur in these numbers for planning purposes.

Program participants submitted biweekly vouchers countersigned by their laboratory research focal point, and RDL issued paychecks so as to arrive in associates' hands two weeks later.

In 1996, RDL implemented direct deposit as a payment option for SFRP and GSRP associates. There were some growing pains. Of the 128 associates who opted for direct deposit, 17 did not check to ensure that their financial institutions could support direct deposit (and they couldn't), and eight associates never did provide RDL with their banks' ABA number (direct deposit bank routing number), so only 103 associates actually participated in the direct deposit program. The remaining associates received their stipend and expense payments via checks sent in the US mail.

HSAP program participants were considered actual RDL employees, and their respective state and federal income tax and Social Security were withheld from their paychecks. By the nature of their independent research, SFRP and GSRP program participants were considered to be consultants or independent contractors. As such, SFRP and GSRP associates were responsible for their own income taxes, Social Security, and insurance.

8. CONTENTS OF THE 1996 REPORT

The complete set of reports for the 1996 SRP includes this program management report (Volume 1) augmented by fifteen volumes of final research reports by the 1996 associates, as indicated below:

1996 SRP Final Report Volume Assignments

LABORATORY	SFRP	GSRP	HSAP
Armstrong	2	7	12
Phillips	3	8	13
Rome	4	9	14
Wright	5A, 5B	10	15
AEDC, ALCs, WHMC	6	11	16

APPENDIX A – PROGRAM STATISTICAL SUMMARY

A. Colleges/Universities Represented

Selected SFRP associates represented 169 different colleges, universities, and institutions, GSRP associates represented 95 different colleges, universities, and institutions.

B. States Represented

SFRP -Applicants came from 47 states plus Washington D.C. and Puerto Rico. Selectees represent 44 states plus Puerto Rico.

GSRP - Applicants came from 44 states and Puerto Rico. Selectees represent 32 states.

HSAP - Applicants came from thirteen states. Selectees represent nine states.

Total Number of Participants	
SFRP	188
GSRP	109
HSAP	138
TOTAL	435

Degrees Represented			
	SFRP	GSRP	TOTAL
Doctoral	184	1	185
Master's	4	48	52
Bachelor's	0	60	60
TOTAL	188	109	297

SFRP Academic Titles	
Assistant Professor	79
Associate Professor	59
Professor	42
Instructor	3
Chairman	0
Visiting Professor	1
Visiting Assoc. Prof.	0
Research Associate	4
TOTAL	188

Source of Learning About the SRP		
Category	Applicants	Selectees
Applied/participated in prior years	28%	34%
Colleague familiar with SRP	19%	16%
Brochure mailed to institution	23%	17%
Contact with Air Force laboratory	17%	23%
IEEE Spectrum	2%	1%
BIIHE	1%	1%
Other source	10%	8%
TOTAL	100%	100%

APPENDIX B -- SRP EVALUATION RESPONSES

1. OVERVIEW

Evaluations were completed and returned to RDL by four groups at the completion of the SRP. The number of respondents in each group is shown below.

Table B-1. Total SRP Evaluations Received

Evaluation Group	Responses
SFRP & GSRPs	275
HSAPs	113
USAF Laboratory Focal Points	84
USAF Laboratory HSAP Mentors	6

All groups indicate unanimous enthusiasm for the SRP experience.

The summarized recommendations for program improvement from both associates and laboratory personnel are listed below:

A. Better preparation on the labs' part prior to associates' arrival (i.e., office space, computer assets, clearly defined scope of work).

B. Faculty Associates suggest higher stipends for SFRP associates.

C. Both HSAP Air Force laboratory mentors and associates would like the summer tour extended from the current 8 weeks to either 10 or 11 weeks; the groups state it takes 4-6 weeks just to get high school students up-to-speed on what's going on at laboratory. (Note: this same argument was used to raise the faculty and graduate student participation time a few years ago.)

2. 1996 USAF LABORATORY FOCAL POINT (LFP) EVALUATION RESPONSES

The summarized results listed below are from the 84 LFP evaluations received.

1. LFP evaluations received and associate preferences:

Table B-2. Air Force LFP Evaluation Responses (By Type)

		How Many Associates Would You Prefer To Get ? (% Response)											
		SFRP				GSRP (w/Univ Professor)				GSRP (w/o Univ Professor)			
Lab	Evals Recv'd	0	1	2	3+	0	1	2	3+	0	1	2	3+
AEDC	0	-	-	-	-	-	-	-	-	-	-	-	-
WHMC	0	-	-	-	-	-	-	-	-	-	-	-	-
AL	7	28	28	28	14	54	14	28	0	86	0	14	0
FJSRL	1	0	100	0	0	100	0	0	0	0	100	0	0
PL	25	40	40	16	4	88	12	0	0	84	12	4	0
RL	5	60	40	0	0	80	10	0	0	100	0	0	0
WL	46	30	43	20	6	78	17	4	0	93	4	2	0
Total	84	32%	50%	13%	5%	80%	11%	6%	0%	73%	23%	4%	0%

LFP Evaluation Summary. The summarized responses, by laboratory, are listed on the following page. LFPs were asked to rate the following questions on a scale from 1 (below average) to 5 (above average).

2. LFPs involved in SRP associate application evaluation process:
    a. Time available for evaluation of applications:
    b. Adequacy of applications for selection process:
3. Value of orientation trips:
4. Length of research tour:
5. a. Benefits of associate's work to laboratory:
    b. Benefits of associate's work to Air Force:
6. a. Enhancement of research qualifications for LFP and staff:
    b. Enhancement of research qualifications for SFRP associate:
    c. Enhancement of research qualifications for GSRP associate:
7. a. Enhancement of knowledge for LFP and staff:
    b. Enhancement of knowledge for SFRP associate:
    c. Enhancement of knowledge for GSRP associate:
8. Value of Air Force and university links:
9. Potential for future collaboration:
10. a. Your working relationship with SFRP:
    b. Your working relationship with GSRP:
11. Expenditure of your time worthwhile:

(Continued on next page)

12. Quality of program literature for associate:
13. a. Quality of RDL's communications with you:
    b. Quality of RDL's communications with associates:
14. Overall assessment of SRP:

Table B-3. Laboratory Focal Point Reponses to above questions

	AEDC	*AL*	*FJSRL*	*PL*	*RL*	*WHMC*	*WL*
# Evals Recv'd	0	7	1	14	5	0	46
Question #							
2	-	86 %	0 %	88 %	80 %	-	85 %
2a	-	4.3	n/a	3.8	4.0	-	3.6
2b	-	4.0	n/a	3.9	4.5	-	4.1
3	-	4.5	n/a	4.3	4.3	-	3.7
4	-	4.1	4.0	4.1	4.2	-	3.9
5a	-	4.3	5.0	4.3	4.6	-	4.4
5b	-	4.5	n/a	4.2	4.6	-	4.3
6a	-	4.5	5.0	4.0	4.4	-	4.3
6b	-	4.3	n/a	4.1	5.0	-	4.4
6c	-	3.7	5.0	3.5	5.0	-	4.3
7a	-	4.7	5.0	4.0	4.4	-	4.3
7b	-	4.3	n/a	4.2	5.0	-	4.4
7c	-	4.0	5.0	3.9	5.0	-	4.3
8	-	4.6	4.0	4.5	4.6	-	4.3
9	-	4.9	5.0	4.4	4.8	-	4.2
10a	-	5.0	n/a	4.6	4.6	-	4.6
10b	-	4.7	5.0	3.9	5.0	-	4.4
11	-	4.6	5.0	4.4	4.8	-	4.4
12	-	4.0	4.0	4.0	4.2	-	3.8
13a	-	3.2	4.0	3.5	3.8	-	3.4
13b	-	3.4	4.0	3.6	4.5	-	3.6
14	-	4.4	5.0	4.4	4.8	-	4.4

3. 1996 SFRP & GSRP EVALUATION RESPONSES

The summarized results listed below are from the 257 SFRP/GSRP evaluations received.

Associates were asked to rate the following questions on a scale from 1 (below average) to 5 (above average) - by Air Force base results and over-all results of the 1996 evaluations are listed after the questions.

1. The match between the laboratories research and your field:
2. Your working relationship with your LFP:
3. Enhancement of your academic qualifications:
4. Enhancement of your research qualifications:
5. Lab readiness for you: LFP, task, plan:
6. Lab readiness for you: equipment, supplies, facilities:
7. Lab resources:
8. Lab research and administrative support:
9. Adequacy of brochure and associate handbook:
10. RDL communications with you:
11. Overall payment procedures:
12. Overall assessment of the SRP:
13. a. Would you apply again?
    b. Will you continue this or related research?
14. Was length of your tour satisfactory?
15. Percentage of associates who experienced difficulties in finding housing:
16. Where did you stay during your SRP tour?
    a. At Home:
    b. With Friend:
    c. On Local Economy:
    d. Base Quarters:
17. Value of orientation visit:
    a. Essential:
    b. Convenient:
    c. Not Worth Cost:
    d. Not Used:

SFRP and GSRP associate's responses are listed in tabular format on the following page.

Table B-4. 1996 SFRP & GSRP Associate Responses to SRP Evaluation

	Arnold	Brooks	Edwards	Eglin	Griffis	Hanscom	Kelly	Kirtland	Lackland	Robins	Tyndall	WPAFB	average
# res	6	48	6	14	31	19	3	32	1	2	10	85	257
1	4.8	4.4	4.6	4.7	4.4	4.9	4.6	4.6	5.0	5.0	4.0	4.7	4.6
2	5.0	4.6	4.1	4.9	4.7	4.7	5.0	4.7	5.0	5.0	4.6	4.8	4.7
3	4.5	4.4	4.0	4.6	4.3	4.2	4.3	4.4	5.0	5.0	4.5	4.3	4.4
4	4.3	4.5	3.8	4.6	4.4	4.4	4.3	4.6	5.0	4.0	4.4	4.5	4.5
5	4.5	4.3	3.3	4.8	4.4	4.5	4.3	4.2	5.0	5.0	3.9	4.4	4.4
6	4.3	4.3	3.7	4.7	4.4	4.5	4.0	3.8	5.0	5.0	3.8	4.2	4.2
7	4.5	4.4	4.2	4.8	4.5	4.3	4.3	4.1	5.0	5.0	4.3	4.3	4.4
8	4.5	4.6	3.0	4.9	4.4	4.3	4.3	4.5	5.0	5.0	4.7	4.5	4.5
9	4.7	4.5	4.7	4.5	4.3	4.5	4.7	4.3	5.0	5.0	4.1	4.5	4.5
10	4.2	4.4	4.7	4.4	4.1	4.1	4.0	4.2	5.0	4.5	3.6	4.4	4.3
11	3.8	4.1	4.5	4.0	3.9	4.1	4.0	4.0	3.0	4.0	3.7	4.0	4.0
12	5.7	4.7	4.3	4.9	4.5	4.9	4.7	4.6	5.0	4.5	4.6	4.5	4.6
Numbers below are percentages													
13a	83	90	83	93	87	75	100	81	100	100	100	86	87
13b	100	89	83	100	94	98	100	94	100	100	100	94	93
14	83	96	100	90	87	80	100	92	100	100	70	84	88
15	17	6	0	33	20	76	33	25	0	100	20	8	39
16a	-	26	17	9	38	23	33	4	-	-	-	30	
16b	100	33	-	40	-	8	-	-	-	-	36	2	
16c	-	41	83	40	62	69	67	96	100	100	64	68	
16d	-	-	-	-	-	-	-	-	-	-	-	0	
17a	-	33	100	17	50	14	67	39	-	50	40	31	35
17b	-	21	-	17	10	14	-	24	-	50	20	16	16
17c	-	-	-	-	10	7	-	-	-	-	-	2	3
17d	100	46	-	66	30	69	33	37	100	-	40	51	46

4. 1996 USAF LABORATORY HSAP MENTOR EVALUATION RESPONSES

Not enough evaluations received (5 total) from Mentors to do useful summary.

5. 1996 HSAP EVALUATION RESPONSES

The summarized results listed below are from the 113 HSAP evaluations received.

HSAP apprentices were asked to rate the following questions on a scale from 1 (below average) to 5 (above average)

1. Your influence on selection of topic/type of work.
2. Working relationship with mentor, other lab scientists.
3. Enhancement of your academic qualifications.
4. Technically challenging work.
5. Lab readiness for you: mentor, task, work plan, equipment.
6. Influence on your career.
7. Increased interest in math/science.
8. Lab research & administrative support.
9. Adequacy of RDL's Apprentice Handbook and administrative materials.
10. Responsiveness of RDL communications.
11. Overall payment procedures.
12. Overall assessment of SRP value to you.
13. Would you apply again next year? Yes (92 %)
14. Will you pursue future studies related to this research? Yes (68 %)
15. Was Tour length satisfactory? Yes (82 %)

	Arnold	Brooks	Edwards	Eglin	Griffiss	Hanscom	Kirtland	Tyndall	WPAFB	Totals
# resp	5	19	7	15	13	2	7	5	40	113
1	2.8	3.3	3.4	3.5	3.4	4.0	3.2	3.6	3.6	3.4
2	4.4	4.6	4.5	4.8	4.6	4.0	4.4	4.0	4.6	4.6
3	4.0	4.2	4.1	4.3	4.5	5.0	4.3	4.6	4.4	4.4
4	3.6	3.9	4.0	4.5	4.2	5.0	4.6	3.8	4.3	4.2
5	4.4	4.1	3.7	4.5	4.1	3.0	3.9	3.6	3.9	4.0
6	3.2	3.6	3.6	4.1	3.8	5.0	3.3	3.8	3.6	3.7
7	2.8	4.1	4.0	3.9	3.9	5.0	3.6	4.0	4.0	3.9
8	3.8	4.1	4.0	4.3	4.0	4.0	4.3	3.8	4.3	4.2
9	4.4	3.6	4.1	4.1	3.5	4.0	3.9	4.0	3.7	3.8
10	4.0	3.8	4.1	3.7	4.1	4.0	3.9	2.4	3.8	3.8
11	4.2	4.2	3.7	3.9	3.8	3.0	3.7	2.6	3.7	3.8
12	4.0	4.5	4.9	4.6	4.6	5.0	4.6	4.2	4.3	4.5
Numbers below are percentages										
13	60%	95%	100%	100%	85%	100%	100%	100%	90%	92%
14	20%	80%	71%	80%	54%	100%	71%	80%	65%	68%
15	100%	70%	71%	100%	100%	50%	86%	60%	80%	82%

A STUDY OF QUANTUM WELLS FORMED IN $Al_xGa_{1-x}As_ySb_{1-y}/In_zGa_{1-z}As/Al_xGa_{1-x}As_ySb_{1-y}$ HETEROSTRCUTURES

A. F. M. Anwar
Associate Professor
Electrical and Systems Engineering Department
The University of Connecticut
Storrs, CT 06269-2157

Final Report for:
Summer Faculty Research Program
Rome Labortory

Sponsored by:
Air Force of Scientific Research
Bolling Air Force Base, DC
and
Rome Laboratory
Hanscom Air Force Base

September 1996

A STUDY OF QUANTUM WELLS FORMED IN $Al_xGa_{1-x}As_ySb_{1-y}/In_zGa_{1-z}As/Al_xGa_{1-x}As_ySb_{1-y}$ HETEROSTRCUTURES

A. F. M. Anwar
Associate Professor
Electrical and Systems Engineering Department
The University of Connecticut
Storrs, CT 06269-2157

Abstract

AlGaAsSb/InGaAs/AlGaAsSb quantum wells are investigated for possible applications in ultra low noise high electron mobility transistors (HEMTs) operating at millimeter wavelength. Schrödinger and Poisson equations are solved self-consistently to calculate the quantum mechanical properties of *AlGaAsSb/InGaAs/AlGaAsSb* single quantum wells formed in HEMTs. The two dimensional electron gas (2DEG) distribution is calculated and shows excellent confinement both at room temperature and at 77K. The variation of the average distance of the electron cloud, from the first heterointerface, with the 2DEG concentration is a strong function of the quantum well (QW) width. A minimum 2DEG concentration threshold, dictated by the QW width and the unintentional doping level of the substrate, exists at room temperature. This effect may prohibit the pinching-off of the channel at room temperature, especially for wide QWs. The room temperature pinch-off properties are strongly affected by the *Al* mole fraction in the buffer layer, the *In* mole fraction in the channel and the unintentional doping level of the lattice matched quaternary buffer. A higher *Al* mole fraction in the buffer along with a lower *In* mole fraction in the channel results in superior pinch-off characteristic. The use of *InAs* as channel material imposes stricter condtions on the composition and the unintentional doping of the buffer layer, while with decreasing *In* mole fraction the restriction is relaxed. Care must be taken to properly choose the *Al* and *In* mole fractions, inability to do that may result in a type-II broken-band or staggered-band configuration.

A STUDY OF QUANTUM WELLS FORMED IN $Al_xGa_{1-x}As_ySb_{1-y}/In_zGa_{1-z}As/Al_xGa_{1-x}As_ySb_{1-y}$ HETEROSTRCUTURES

A. F. M. Anwar

1. Introduction

InAs, with its low effective mass, high low field mobility and a large Γ-L separation is an ideal channel material for high electron mobility transistors (HEMTs) [1,2]. These properties make *InAs* HEMTs strong candidates to lower amplifier noise figures and extend operating frequencies to the sub-millimeter band. A host of binary, ternary and quaternary barrier materials are being investigated to create a HEMT structure which preserves the desirable transport properties while maximizing the number of carriers confined to the channel [3–5]. *InAs/AlSb*, a strained layer type -II material system with a conduction band discontinuity $\Delta E_c = 1.35$eV, suffers from defects that may lead to kinks in the I-V characteristics [5]. The ternary *AlAsSb*, which is lattice matched to *InAs* and offers a large ΔE_c, suffers from a lower low field mobility in *InAs* [6]. The quaternary $Al_xGa_{1-x}As_ySb_{1-y}/InAs$, also latticed matched to *InAs*, offers both a higher low field mobility in *InAs* and a large ΔE_c.

However, for such deep quantum wells (QW) there is a possibility that the 2DEG concentration can not be reduced below a certain number and may lead to the inability to pinch-off the channel [6]. Similar problems are not observeable in *AlGaAs/GaAs* or *InGaAs/InAlAs* heterostructures as the conduction band discontinuity is always much smaller then the band gap of the buffer layer.

The report is organized as follows. A general introduction is presented in Section 1. In Section 2, a method to determine the bandgap of the quaternary *AlGaAsSb* and the conduction band discontinuity at the *AlGaAsSb/InGaAs* heterointerface is presented. In Section 3, a quasi-analytical model is described to solve Schrödinger and Poisson's equations, self-consistently, for deep quantum wells formed in *AlGaAsSb/InGaAs/AlGaAsSb* systems. The effect of the presence of deep quantum wells on the pinch-off characteristics of HEMTs is discussed in Section 4. Sections 2 through 4, end with a conclusion.

2.1 Bandgap of the Quaternary and the Conduction Band Discontinuity in $AlGaAsSb/InGaAs$ Heterointerfaces

The basic parameters required to model any quantum structure are the band discontinuities and the bandgaps of the constituent materials. Experimental bandgap data for the quaternary $AlGaAsSb$ are extremely limited. Data on band discontinuities formed by the quaternary and other binary/ternary compounds are rare or are non-existent. In this section, for the first time, the bandgap of the quaternary and band discontinuities formed with $InGaAs$ is presented [7].

2.2 Approach

Our method for calculating the conduction band discontinuity ΔE_c is an extension of the approach reported by Schuermeyer *et. al.*[8], and is illustrated graphically in Fig.1. The diagram is constructed beginning with known valence band energy differences for binary systems and then adding the bandgap energies to find the conduction band minimum for each binary compound. The energy bandgaps of the ternary systems are computed by interpolation over alloy composition using bowing parameters [7,9,10,11]. The calculation of the energy bandgap and the lattice constants of the quaternary follows the work of Moon *et. al.* [9], Glisson *et. al.* [10] and Svennson *et. al.* [11]. The energy bandgap of the quaternary can be written as:

$$E_G^Q(x,y) = (1-x)T_{14}(y) + xT_{23}(y) - \Delta(x,y) \tag{1}$$

where $T_{ij}(y) = yB_j + (1-y)B_i - y(1-y)C_{ij}$ are the ternary alloy bandgaps, B_i's are the bandgap of the binaries, C_{ij} are the bowing parameters for the ternary alloy, and $\Delta(x,y) = x(1-x)[(1-y)C_{12} + yC_{43}] + x(1-x)y(1-y)C_Q$, where $C_Q = \frac{C_{14}}{x} + \frac{C_{23}}{1-x}$. Using the parameters listed in Table-1[11], $E_G^Q(x,y)$ is calculated for the Γ-, X-, and L-valleys and the lowest result is chosen as the bandgap. The lattice constant of the quaternary L_Q is interpolated using the following relationship:

$$L_Q = L_1 + (L_2 - L_1)x + (L_4 - L_1)y + (L_1 - L_2 + L_3 - L_4)xy \tag{2}$$

where L_i's are the lattice constants of the binaries listed in Table-1[11]. Finally, ΔE_c for the lattice matched $Al_xGa_{1-x}As_{1-y}Sb_y/InAs$ is determined by calculating the difference between the conduction band energies of $InAs$ and $Al_xGa_{1-x}As_{1-y}Sb_y$ as shown in Fig.1.

2.3 Results and Discussions

Fig.2 shows the energy bandgap of $Al_xGa_{1-x}As_{1-y}Sb_y$ lattice matched to $InAs$ as a function of Al mole fraction, x. On the same graph the conduction band discontinuity, ΔE_c, of $Al_xGa_{1-x}As_{1-y}Sb_y/InAs$ is plotted. The ΔE_c is linear with x. For $x < 0.2$, $E_G^Q < \Delta E_c$, resulting in the type-II broken-band alignment[8] illustrated by the inset in Fig. 2. For $x > 0.6$, $E_G^Q \geq \Delta E_c + E_G^{channel}$, where $E_G^{channel}$ is the bandgap of the channel material, giving a type-I alignment[12]. Between $x = 0.2$ and $x = 0.6$, $\Delta E_c + E_G^{channel} > E_G^Q > \Delta E_c$, and the band structure is type-II staggered-band[12].

In Fig.3, the bandgap of the quaternary is plotted as a function of the Sb mole fraction with Al mole fraction, x, as a parameter. The Γ-valley has the lowest conduction band energy for alloys below the shaded region of the plot. These alloys have a direct bandgap. In the shaded region, the L-valley determines the bandgap. Above the shaded region, the X-valley has the lowest energy. For alloys in and above the shaded region, the bandgap is always indirect. Thus, the change from a direct to an indirect bandgap material is from Γ to L followed by a change from L to X. With increasing y the change from L to X moves from $x = 0.65$ at $y = 0$ to $x = 0.55$ at $y = 1$. Near $x = 0.6$ and $y = 0.65$, the Γ- and L-valley energies are equal, reducing the extent of the L-valley region to a point. A change directly from Γ to X may be possible. Also shown on the plot are the ranges of x and y for lattice matched conditions to be satisfied. For $AlGaAsSb$ to be lattice matched to $InAs$ a very narrow window of Sb mole fraction ranging from 0.83 to 0.92 is allowed. For lattice matched $AlGaAsSb/In_{0.8}Ga_{0.2}As$ and $AlGaAsSb/In_{0.52}Ga_{0.42}As$ the Sb mole fraction range is 0.67-0.73 and 0.43-0.48, respectively. The lattice matching condition can always be satisfied for any value of x.

In Fig.4, the bandgaps of the quaternaries and ΔE_c of the lattice matched $Al_xGa_{1-x}As_{1-y}Sb_y/In_{0.8}Ga_{0.2}As$ and $Al_xGa_{1-x}As_{1-y}Sb_y/In_{0.52}Ga_{0.48}As$ are plotted as function of Al mole fraction x (also see Table-2 and Table-3). The bandgap of the quaternary changes from the Γ- to the X-valley for the $AlGaAsSb/In_{0.8}Ga_{0.2}As$ system near $x = 0.6$. A crossover from Γ to L near $x = 0.6$ is followed by a change from L to X near $x = 0.7$ for the $AlGaAsSb/In_{0.52}a_{0.48}As$ system. These quaternary/ternary heterostructures do not show a type-II broken-band alignment. A change from type-II staggered to type-I takes place for

the $In_{0.8}Ga_{0.2}As$ and $In_{.52}Ga_{.48}As$ heterostructures at Al mole fractions of 0.55 and 0.5, respectively.

2.4 Conclusion

In conclusion, bandgaps of the quaternary $AlGaAsSb$ lattice matched to $InAs$ and $InGaAs$ and the corresponding conduction band discontinuities are presented for varying Al and Sb mole fractions. A change from type-II broken-band to type-II staggered band alignment is observed for lattice matched $AlGaAsSb/InAs$ systems. Type-II broken band alignments are absent in the $AlGaAsSb/In_{0.8}Ga_{0.2}As$ and $AlGaAsSb/In_{.52}Ga_{.48}As$ lattice matched systems. The minimum Al mole fraction required to obtain a type-I band alignment increases with increasing In mole fraction. The conduction band discontinuity increases with In concentration. The largest conduction band discontinununity is obtained in the lattice matched $AlGaAsSb/InAs$ material system.

3.1 An Envelope Function Description of Deep Quantum Wells

In this section we report a self-consistent solution to model the QW formed in the conduction band of an $AlGaAsSb/InAs/AlGaAsSb$ heterostructure [13]. The results of this analysis will enable us to compute the charge control, current-voltage and noise performance of this class of devices. Furthermore, this analysis will guide material, device and fabrication research to achieve ultra low noise, very high frequency HEMTs.

3.2 Mathematical Model

In $AlGaAsSb/InAs/AlGaAsSb$ systems the QW is formed in InAs (see Fig. 5). Following the method previously developed by the author (Ref. 14), the one electron Schrödinger equation, under effective mass approximation, can be written as

$$-(\hbar^2/2)\frac{d}{dx}(\frac{1}{m^*}\frac{d\zeta_i}{dx}) + (V(x) - E_i)\zeta_i = 0 \qquad (3)$$

where m^* is the electron effective mass, $\hbar$ is the reduced Planck's constant, $\zeta_i(x)$ is the envelope wave function, E_i is the energy eigen value, $V(x)$ is the potential energy and the subscript i denotes the ith subband. For simplicity the potential energy function is approximated by three straight lines with slopes a_1, a_2 and a_3, respectively, and is expressed as

$$V(x) = \begin{cases} V_0 & x < 0 \\ a_j x + \Delta E_j & x_{j-1} < x < x_j \quad \text{j=1,2,3} \end{cases}$$

$\Delta E_3 = \Delta E_2 + \Delta E_{c2} + L(a_2 - a_3)$, ΔE_{c2} is the second heterointerface, $L = x_2 - x_0$ is the width of first heterointerface, and x_3 is the distance from the the electrons reside. The solution to the Schrödinger be written as

$$\begin{matrix} _{,0}e^{\beta x} + \alpha_{2,0}e^{-\beta x} & j = 0 \\ _{1,j}Ai(\xi_j) + \alpha_{2,j}Bi(\xi_j) & j = 1,2,3 \end{matrix} \quad (6)$$

determined by Poisson's
can be expressed in

$\overline{E}$), ΔE_{c1} is the conduction band discontinuity at the first
on effective mass in $AlGaAsSb$, Ai and Bi are the Airy and
heterointerface, $\pi\iota_0$ --
the complementary function, respectively, $\xi_j = \gamma_j(x + \frac{\Delta E_j - E}{a_j})$, with $\gamma_j = (2m_j^* a_j)/\hbar^2)^{1/3}$ and $\alpha_{k,j}$'s ($k = 1,2$ & $j = 0,1,2,3$) are the arbitrary constants. Here the subscript j refers to region j and the superscript i, whenever used, will refer to the *ith* subband. The eigen values and eigen functions are determined by applying the two boundary conditions at any interface (a) continuity of the wave function and (b) continuity of the first derivative by taking into account the proper effective mass.

Having formulated the Schrödinger equation Poisson's equation is formulated:

$$\epsilon\frac{d^2\phi}{dx^2} = q\sum_i n_{si}\zeta_i^2(x) + qN_A \quad (4)$$

where $n_{si} = \frac{m^*_{InAs}kT}{\pi\hbar^2} ln(1 + e^{(E_F - E_i)/kT})$ is the number of electrons per unit area, $\zeta_i(x)$ is the envelope wave function in the ith subband, N_A is the acceptor density in the unintentionally doped $AlGaAsSb$ layer, m^*_{InAs} refers to the effective mass in the channel, T is the temperature and E_F is the Fermi level at the interface relative to the conduction band in the channel at $x = 0$. In these equations we have chosen the potential energy at the interface as the reference. The Fermi level E_F is expressed as

$$E_F = q[\phi(0) - \phi(W)] + E_{F0} + \Delta E_{c2} \quad (5)$$

where $\phi(0) - \phi(W)$ is the total band bending in the $AlGaAsSb$ layer $\phi(0) = 0$, W is the depletion depth, $E_{F0} = -\left[\frac{E_g(T)}{2} + kTln\frac{N_A}{n_i(T)}\right]$ is the position of the Fermi level with respect to the conduction band in the bulk $AlGaAsSb$ and $n_i(T)$ is the intrinsic carrier concentration in $AlGaAsSb$. The slopes a_j of the straight lines, which approximate the

shape of the QW, are proportional to the average electric field d
equation. By integrating eqn. (2) twice with respect to x, the slopes c
the form

$$a_j = (q^2/\epsilon)(f_j n_s + N_A W), \quad j = 1, 2, 3$$

where

$$f_j = 1 - \sum_i \frac{n_{si}}{n_s} \frac{1}{x_j - x_{j-1}} \int_{x_{j-1}}^{x_j} dx \int_{-\infty}^{x} \zeta_i^2(x')dx', \quad j = 1, 2, 3 \tag{7}$$

and $n_s = \sum_i n_{si}$ is the channel electron density in cm^{-2}.

By solving the one electron Schrödinger equation for the given potential we can obtain the eigen energies and the wave functions for the system. The eigen energies and the wave functions determine the shape of the electron distribution in the quantum well which is then used to solve Poisson's equation. The two equations are solved self-consistently until we have accounted for 99% of the carriers in the quantum well.

3.3 Results and Discussions

The calculations in this section are based upon the following data for $Al_{0.65}Ga_{0.35}As_{0.1}Sb_{0.9}/InAs$: (a) conduction band discontinuity $\Delta E_{c1} = \Delta E_{c2} = \Delta E_c = 1.35\text{eV}$ [2] (b) $m^*_{InAs} = 0.0239m_0$ (c) $m^*_{AlGaAsSb} = 0.0955m_0$ [11], where m_0 is the free electron mass. Using a band gap of 1.549 eV [11], the intrinsic carrier concentration in the buffer layer is estimated to be $2 \times 10^5\ cm^{-3}$. The unintentional doping of the buffer layer is assumed to be $1 \times 10^{15}\ cm^{-3}$ p-type.

In Fig. 5, the calculated conduction band profile for a QW width of 150Å is shown along with the 2DEG distribution. The plots are obtained for $n_s = 2.0 \times 10^{12}\ cm^{-2}$. Due to the occupancy of higher sub-bands the 2DEG distribution shows a slight hump around 100Å. In the 2DEG distribution for QW widths of 100Å and 50Å this hump is absent. The hump disappears because in narrow QWs (50Å and 100Å) the energy levels are farther apart reducing the occupancy of the sub-band. The sub-band occupancy at both 77K and 300K are plotted in Fig. 6. As is evident, carrier confinement gets better with decreasing temperature.

In Fig. 7, the 2DEG concentration is plotted as a function of the position of the Fermi

level above the bottom of the QW. The concentration is plotted for both 77K and 300K with QW width as a parameter. At 77K, as the bottom of the well is raised toward the Fermi level, the concentration decreases slowly at first, then more rapidly below 0.4 eV. This faster rate of decrease is due to the change in shape of the QW from a rectangular to a more triangular shape. The same behavior is observed for all the QW widths. However, at 300K the eigen energies for wider quantum wells are very close to the Fermi level which is established by the unintentional doping of the layer beneath the channel. The separation of the first eigen energy and the Fermi level for 150Å QW is 0.04eV, whereas, for a 50Å QW this separation is 0.2eV. This condition causes the 2DEG concentration to stop decreasing when the Fermi level reaches approximately 0 eV at room temperature. For wider QWs, the minimum 2DEG concentration decreases with increasing doping level of the unintentionally doped buffer layer as shown in Fig. 8. This observation leads to the important conclusion that pinching-off a deep QW channel at room temperature will be extremely difficult as the 2DEG concentration is always finite for practical gate bias voltages and low doping level of the underlying buffer layer. Failure to pinch-off, in a similar deep QW, was experimentally observed by Li *et. al.* [6]. At 77K, or for narrower QWs, the Fermi level is well below the eigen energies of the QW, therefore, the 2DEG concentration can decrease indefinitely, making pinch-off possible.

In Fig. 9, x_{av} is plotted as a function of n_s. x_{av} for both 77K and 300K starts around $L/2$ at very low concentration, reflecting the presence of a deep rectangular QW. For wider QWs the higher sub-bands are easily populated (see Fig. 6) and ,therefore, the variation in x_{av} over the range of n_s is large. From the figure, for the 50Å QW x_{av} changes approximately 3Å over an n_s variation from 1×10^6 cm^{-2} to 1.2×10^{13} cm^{-2}, whereas, for the 150Å QW the corresponding change in x_{av} is more than 20Å.

3.4 Conclusion

In conclusion, a self-consistent model to investigate the QW formed in $Al_xGa_{1-x}As_ySb_{1-y}$ is presented. A large ΔE_c provides excellent carrier confinement both at 300K and 77K. The unintentional doping density of the higher band-gap substrate plays a major role in establishing the minimum 2DEG concentration in the channel. This constraint on the 2DEG concentration is severe for a wider QW (150Å) and may lead to the absence of any

pinch-off in the current-voltage characteristics at room temperature.

4.1 On The Possible Effects of AlGaAsSb Growth Parameters On The 2DEG Concentration in AlGaAsSb/InGaAs/AlGaAsSb QWs

By introducing Ga in the channel layer, the properties of the QW can be tailored to maintain the desired transport properties while eliminating the pinch-off problem. In $AlGaAsSb/InGaAs/AlGaAsSb$ systems the conduction band discontinuity is comparable to the band gap of the buffer and the barrier layer and results in some interesting quantum properties that may be attributed only to deep QWs. The large conduction band discontinuity obtainable in a $Al_xGa_{1-x}As_{1-y}Sb_y/In_zGa_{1-z}As$ heterostructure depends to a large extent upon the mole fractions x, y and z [7]. The band gap, E_G, also is a strong function of x and y for a given z and plays a major role in the calculation of the properties of the QW formed in $In_zGa_{1-z}As$. An additional constraint is imposed on x and y by the need to match the lattice constants of the buffer, channel and barrier layers. In this section, the effect of process/design parameters on the pinch-off performance of an HEMT is addressed [15].

4.2 Results and Discussions

The results reported are based upon a self-consistent solution of Schrödinger and Poisson's eqations [14,15]. The lowest band gap and the electron/hole effective mass of the quaternary is determined by following the method of Moon *et. al.* [7,9]. The lowest band gap, depending upon the Al and Sb mole fractions can either be direct or indirect. In Fig.10, the lowest energy band gaps along with $\Delta E_c/E_G$ is plotted as a function of the Al mole fraction. $\Delta E_c/E_G$ is greater than 1 for $InAs$ for $x < 0.25$ and the $AlGaAsSb/InAs$ system has a type-II band alignment that eventually changes to a type-I for $x > 0.25$. $In_{0.8}Ga_{0.2}As$ also has a type-II band alignment [7] for very small x. $In_{0.52}Ga_{0.48}As$ is always type-I in a $AlGaAsSb/In_{0.52}Ga_{0.48}As$ latticed matched system. For $InAs$ and $In_{0.52}Ga_{0.48}As$, the band structure changes from Γ to L (at $x = 0.4$ for $InAs$ and $x = 0.6$ for $In_{0.52}Ga_{0.48}As$) and then from L to X (at $x = 0.6$ for $InAs$ and $x = 0.7$ for $In_{0.52}Ga_{0.48}As$) with increasing Al mole fraction. For $In_{0.8}Ga_{0.2}As$ the band structure changes from Γ to X around x=0.6. Electron effective mass in the buffer layer, $m_Q = m_{GaAs} + (m_{AlAs} - m_{GaAs}) * x + (m_{GaSb} - m_{GaAs}) * y + (m_{GaAs} - m_{AlAs} + m_{AlSb} - m_{GaSb}) * x * y$ where x and y are the Al and

Sb mole fractions, respectively. $m_{GaS} = 0.067m_0$, $m_{AlAs} = 0.15m_0$, $m_{AlSb} = 0.12m_0$ and $m_{GaSb} = 0.042m_0$, where m_0 is the free electron mass. The calculation of the conduction band discontinuity follows the method suggested by the present author [7] and is elaborated in Section 2..

In Fig. 11, the minimum 2DEG concentration in the channel,$n_{2D,min}$ is plotted as a function of the unintentional doping level of the AlGaAsSb buffer layer, N_A, with the QW width as a parameter for Al mole fractions of 0.6 and 0.8, respectively. The plots are obtained by assuming a negative gate bias which provides nearly flat band conditions in the AlGaAsSb buffer layer. This condition also yields $n_{2D,min}$ in the $InAs$ channel. Although even lower $n_{2D,min}$ may be obtained by applying an abnormally large negative gate bias, it would create an accumulation of holes at the $InAs$/buffer layer interface, and would prove to be detrimental to HEMT operaton. As shown in Fig.11, $n_{2D,min}$ decreases with N_A and increasing Al mole fraction. Morever, $n_{2D,min}$ decreases with decreasing QW width and is quite negligible for QW widths less than 50Å.

The effect of varying Al mole fraction in the buffer layer on $n_{2D,min}$ is shown in Fig. 12. The plots are obtained for a QW width of 150Å and for $N_A = 1\times10^{14}$, 5×10^{14} and 1×10^{15} cm^{-3}. In mole fraction z = 1, 0.8 and 0.52 are considered. The $AlGaAsSb/InGaAs$ is lattice matched. As observed, irrespective of the channel material $n_{2D,min}$ decreases with x. However, the decrease in $n_{2D,min}$ with x, for $InAs$, is much slower than that for the other two material systems. Over the range of x considered, $n_{2D,min}$ for $InAs$ $In_{.8}Ga_{.2}As$ and $In_{.52}Ga_{.48}As$ changes by 6, 10 and 14 orders of magnitude, respectively, and is a direct consequence of the beahavior of E_G and the conduction band discontinuity ΔE_c with x. It is to be noted that at lower x $n_{2D,min}$ is appreciable and a nearly flat conduction band in the buffer layer can not be satisfied.

Referring back to Fig.10, E_G increases with decreasing In mole fraction for the same x. On the other hand, ΔE_c decreases with decreasing z for the same x. A lower E_G implies a larger intrinsic carrier concentration. Therefore, for the same N_A the Fermi level will be closer to the valence band in $InAs$ as compared to the lower In mole fraction alloys. The quantity of interest is the difference between the first eigen energy and the Fermi level

$\delta E_{1F} = E_1 - E_F$ [15]. A Fermi level close to the valence band compounded by a large ΔE_c (deep QW) will make it possible for δE_{1F} to be a very small positive or even a negative number resulting in an appreciable $n_{2D,min}$. With increasing N_A, δE_{1F} increases implying a decreasing $n_{2D,min}$ as is already shown in Fig. 11. Moreover, with decreasing QW width, for the same x and N_A , δE_{1F} increases (decreases) for $E_1 > E_F$ ($E_1 < E_F$) resulting in a lower $n_{2D,min}$.

4.3 Conclusion

The interplay between the band gap and the conduction band discontinuity and their dependence upon the *Al*, *Sb* and *In* mole fraction impart some interesting properties to the latticed matched *AlGaAsSb/InGaAs* system. A lower unintentional doping in the buffer layer results in a higher $n_{2D,min}$. The effect of the buffer layer *Al* mole fraction on $n_{2D,min}$ is much more stronger than that of N_A. $n_{2D,min}$ always decreases with decreasing *In* mole fraction. For successful HEMT operation 80% *In* is suggested with higher *Al* concentration. The use of 80% *In* will enable the realization of wider QW HEMTs (150Å) that will impart to it an extremely low noise operation capability.

References

1. K. Yoh, T. Moriuchi and M. Inoue, IEEE Electron Dev. Lett., vol. 11, no. 11, p. 526, 1990.

2. S. Miya, S. Muramatsu, N. Kuze,, K. Nagase, T. Iwabuchi, A. Ichii, M. Ozaki and I. Shibasaki, J. Electron. Mat., vol. 25, no. 3, p. 415, 1996.

3. J. A. Lott, L. R. Dawson, E. D. Jones and J. F. Klem, Appl. Phys. Lett., vol. 56, no. 13, p. 1242, 1990.

4. C. R. Bolognesi, E. J. Caine and H. Kroemer, IEEE EDL-15, p. 16, 1994.

5. J. B. Boos, W. Kruppa, B. V. Shanabrook, D. Park, J. L. Davis and H. B. Dietrich, Electron. Lett., vol. 29, no. 21, p. 1888, 1993

6. X. Li, K. F. Longbeach, Y. Wang and W. I. wang, IEEE EDL vol. 13, no. 4, p.192, 1992.

7. A. F. M. Anwar and R. T. Webster, to be submitted to Appl. Phys. Lett.

8. F. L. Schuermeyer, P. Cook, E. Martinez and J. Tantillo, Appl. Phys. Lett., vol. 55, no. 18, p. 1877, 1989.

9. R. L. Moon, G. A. Antypas and L. W. James, J. Electron. Mat., vol. 3, no. 3, p. 635, 1974.

10. T. H. Glisson, J. R. Hauser, M. A. Littlejohn and C. K. Williams, J. Electron. Mat., vol. 7, no. 1, 1978.

11. S. P. Svensson, D. M. Gill and P. N. Uppal, to be published in J. Appl. Phys.

12. E. T. Yu, J. O. McCaldin and T. C. McGill *Band Offsets in Semiconductor Heterojunctions*, Solid State Physics, Eds. H. Ehrenreich and D. Turnbull (Academic Press, Inc., 1992), vol. 46.

13. A. F. M. Anwar and R. T. Webster, Jour. Appl. Phys., 15th Dec., 1996.

14. A. F. M. Anwar and A. N. Khondker, J. Appl. Phys., vol 62, no. 10, p. 4200, 1987.

15. A. F. M. Anwar and R. T. Webster, to be submitted to IEEE EDL.

Table 1-A: Energy bandgap and lattice constants of the binaries[7]

Binary Alloy	B_Γ	B_X	B_L	L(Å)
GaAs (1)	1.424	1.9	1.708	5.64
AlAs (2)	2.892	2.168	2.35	5.66
AlSb (3)	2.223	1.589	1.879	6.135
GaSb (4)	0.726	1.02	0.799	6.094

Table 1-B: Bowing parameters of the ternary alloys[7]

Bowing Parameter	AlGaAs (1,2)	GaAsSb (1,4)	AlAsSb (2,3)	AlGaSb (4,3)
$C_\Gamma(i,j)$	0.041	1.2	0.72	0.368
$C_X(i,j)$	0.125	0.0	0.0	0.077
$C_L(i,j)$	0.45	0.0	0.0	0.334

Table 2: Compositional Variation of The Energy Gap of the Quaternary and ΔE_c

Material System	E_G	ΔE_c
$Al_xGa_{1-x}As_{1-y}Sb_y/InAs$	$0.676 + 1.2489x - 0.254x^2$	$0.8565 + 0.255x$
$Al_xGa_{1-x}As_{1-y}Sb_y/In_{0.8}Ga_{0.2}As$	$0.644 + 1.226x - 0.08x^2$	$0.686 + 0.279x$
$Al_xGa_{1-x}As_{1-y}Sb_y/In_{.52}Ga_{.48}As$	$0.754 + 1.26x - 0.086x^2$	$0.445 + 0.304x$

Table 3: ΔE_c-E_G Relationship

Material System	$\Delta E_c(E_G)$
$Al_xGa_{1-x}As_{1-y}Sb_y/InAs$	$0.67 + 0.26\ E_G$
$Al_xGa_{1-x}As_{1-y}Sb_y/In_{0.8}Ga_{0.2}As$	$0.527 + 0.24\ E_G$
$Al_xGa_{1-x}As_{1-y}Sb_y/In_{.52}Ga_{.48}As$	$0.25 + 0.26\ E_G$

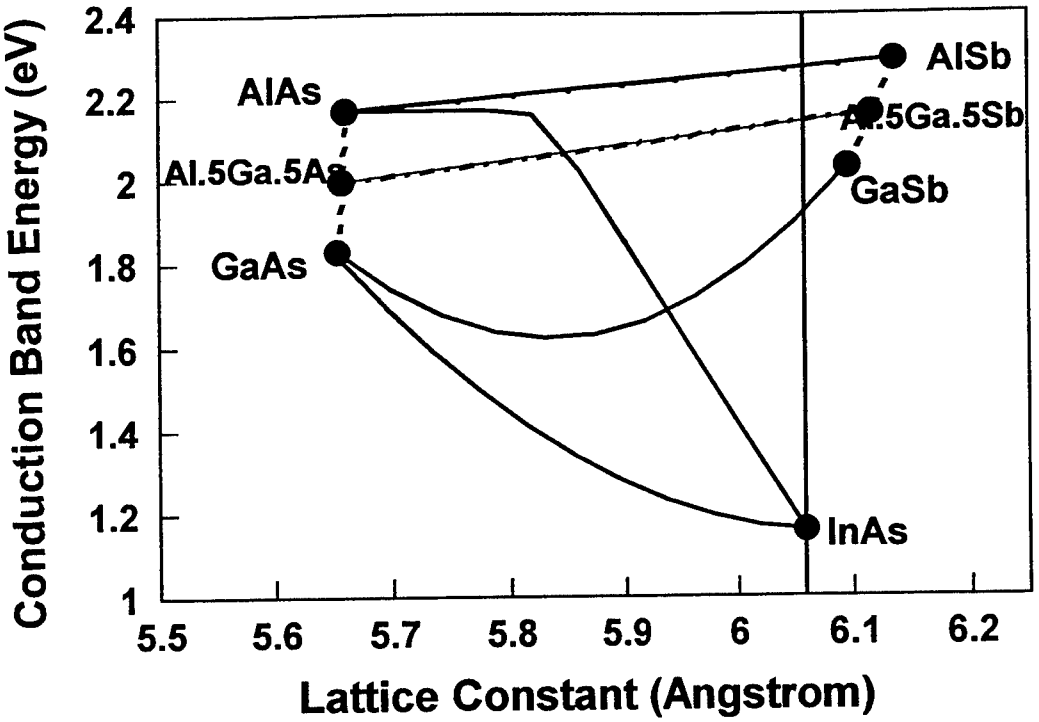


Fig.1 Conduction band energy of the ternary compounds, referenced to the valence band of AlAs, are plotted as a function of the lattice constant. The conduction band energy of the quaternary $Al_{.5}Ga_{.5}As_{1-y}Sb_y$ is shown using the dot-dashed line. The intersection of this line with the InAs line gives the conduction band energy of $Al_{.5}Ga_{.5}As_{1-y}Sb_y$ lattice matched to InAs.

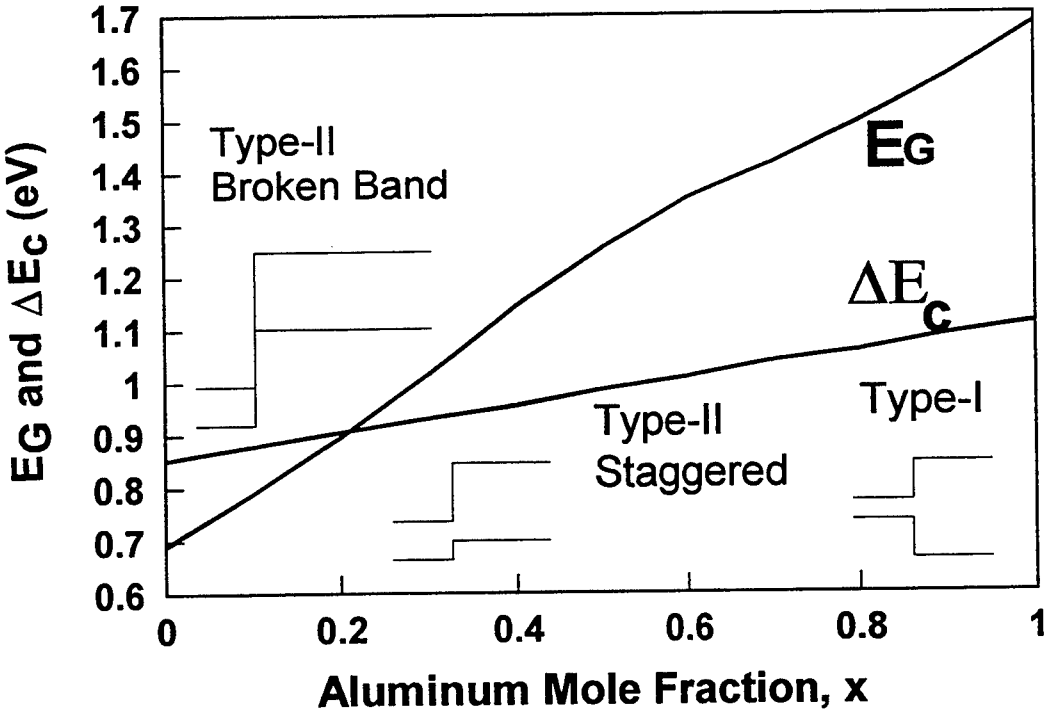


Fig.2 Energy bandgap (solid line) and the conduction band discontinuity (dashed line) of $Al_xGa_{1-x}As_{1-y}Sb_y$ lattice matched to InAs vs. Al mole fraction with type-I and type-II bands as inset.

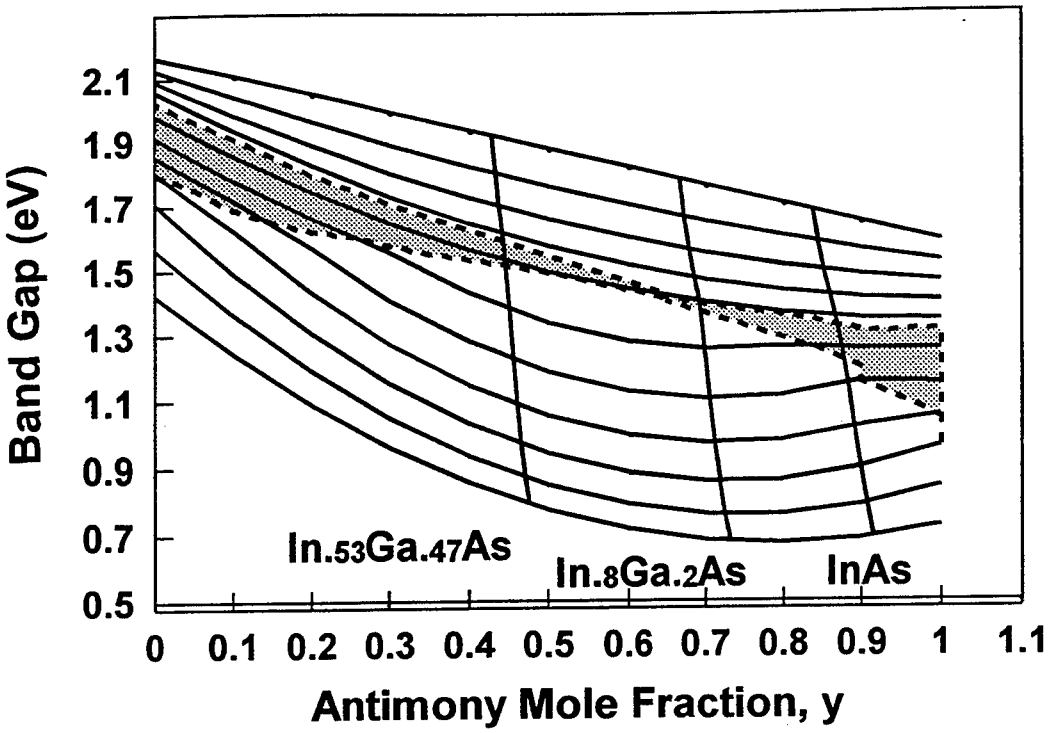


Fig.3 Energy bandgap of Al_xGa_{1-x} As_{1-y} Sb_y as a function of Sb mole fraction with Al mole fraction as a parameter. The lowest bandgap Γ (below the shaded region), L (shaded region) and X (above the shaded region).

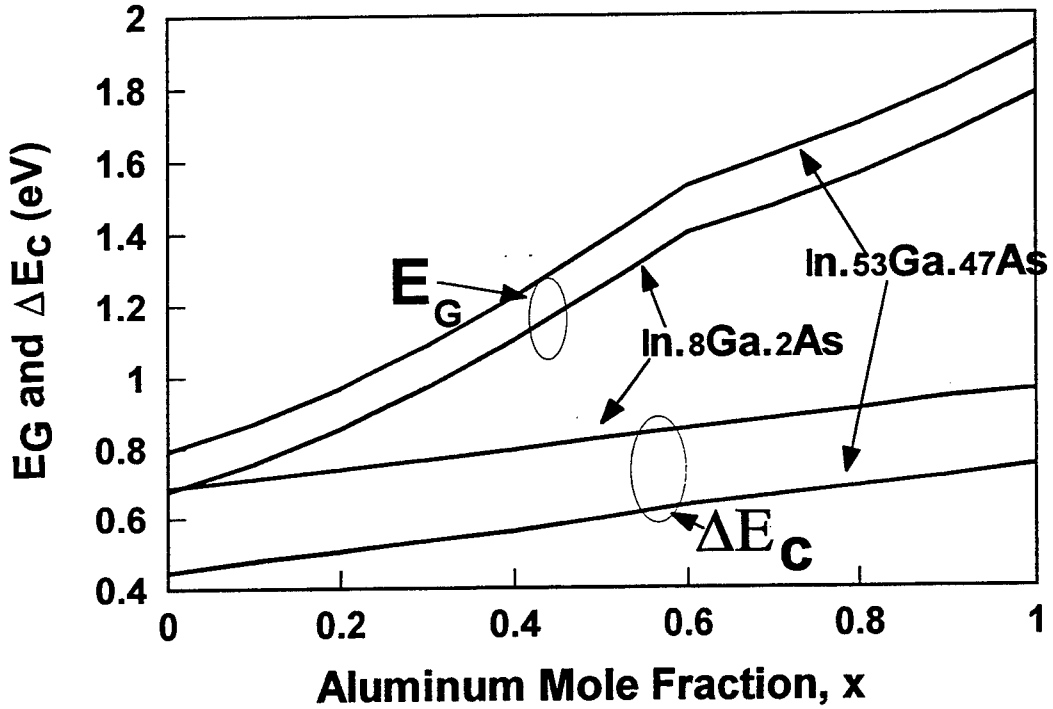


Fig.4 E_G of the quaternary and ΔE_c of the lattice matched Al_xGa_{1-x} As_{1-y} Sb_y /$In_{.8}$ $Ga_{.2}$ As and Al_xGa_{1-x} As_{1-y} Sb_y /$In_{.52}$ $Ga_{.48}$ As vs. x.

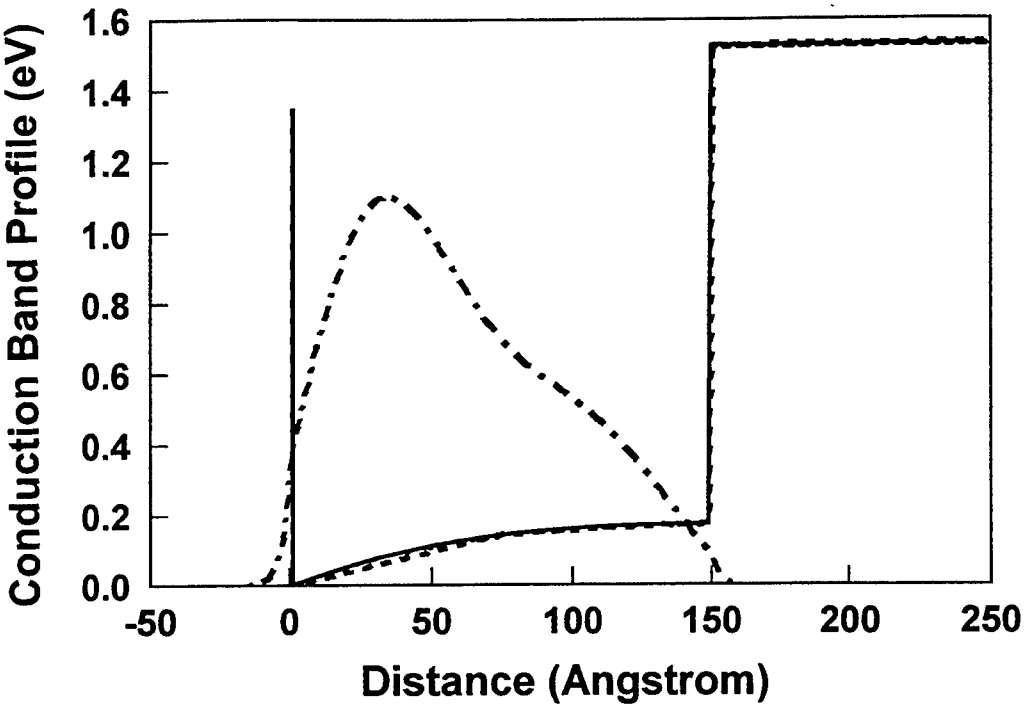


Fig.5 Conduction band profile and 2DEG concentration are plotted as a function of distance for 300K. An unintentional doping of $1x10^{15}$ cm^{-3} is assumed. The plot is made for $n_s = 2x10^{12}$ cm^{-2} and QW width L=150A.

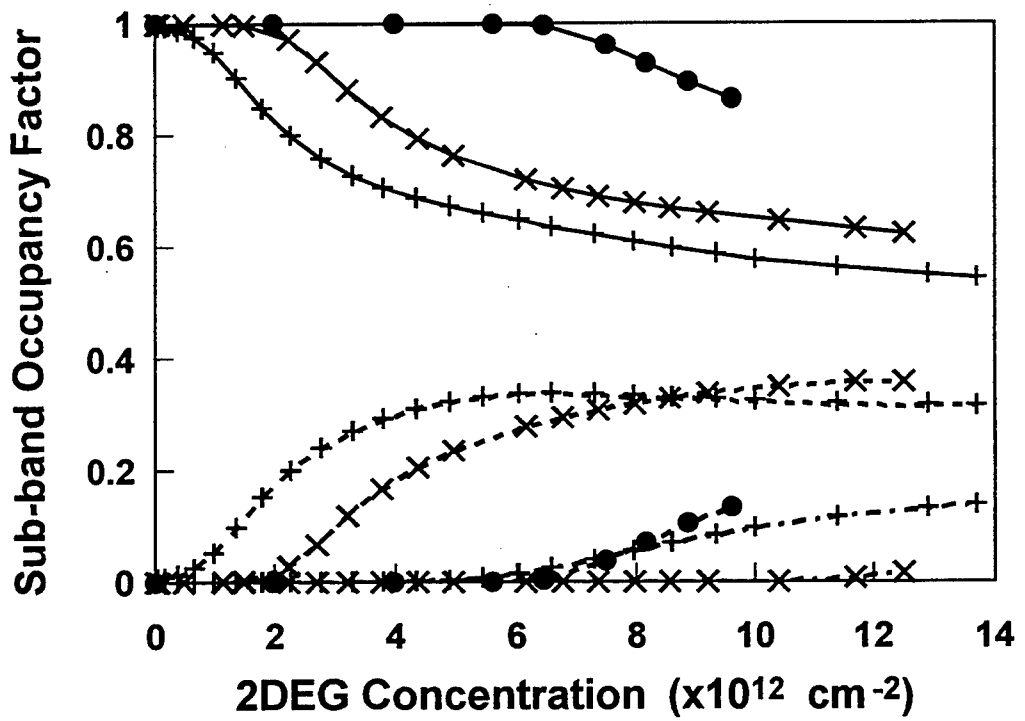


Fig. 6 Sub-band occupancy factors f_i are plotted as a function of n_s for the first (solid line), second (long-dash) and third (short-dash) eigen-energies, respectively, at 300K. The well width considered are, 150A (+), 100A (x) and 50A (•).

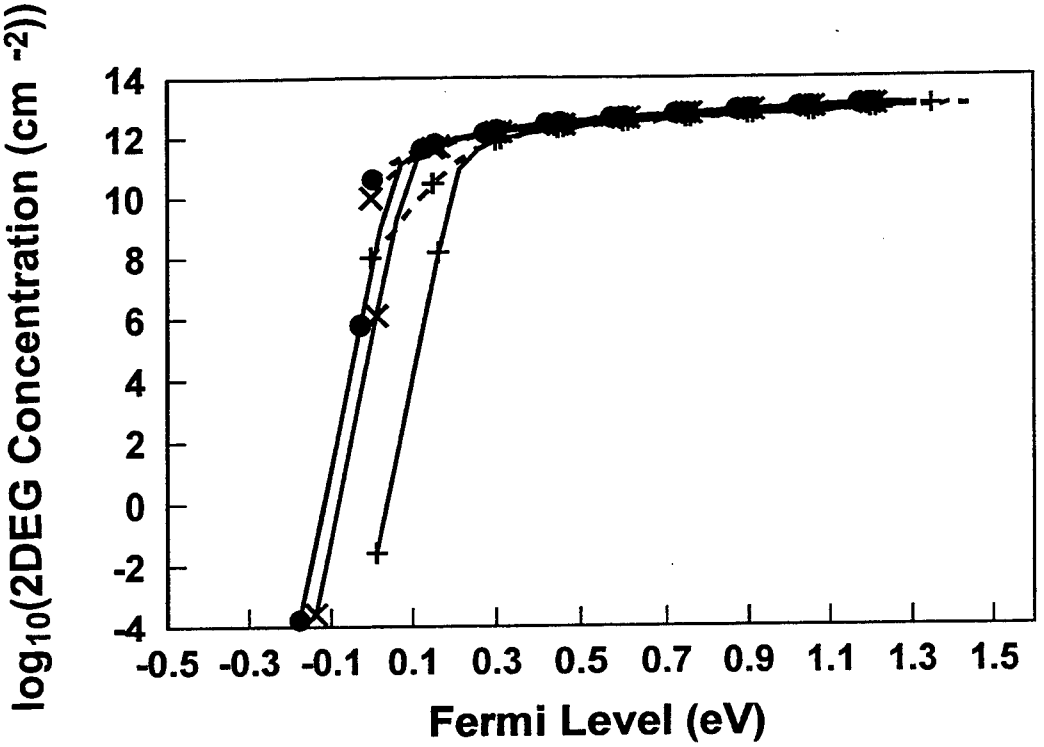


Fig.7 The position of the Fermi level E_F is plotted as a function of n_n for well widths of 50A (+), 100A (x) and 150 A (•), respectively. The solid lines represent 77K and the dashed lines represent 300K.

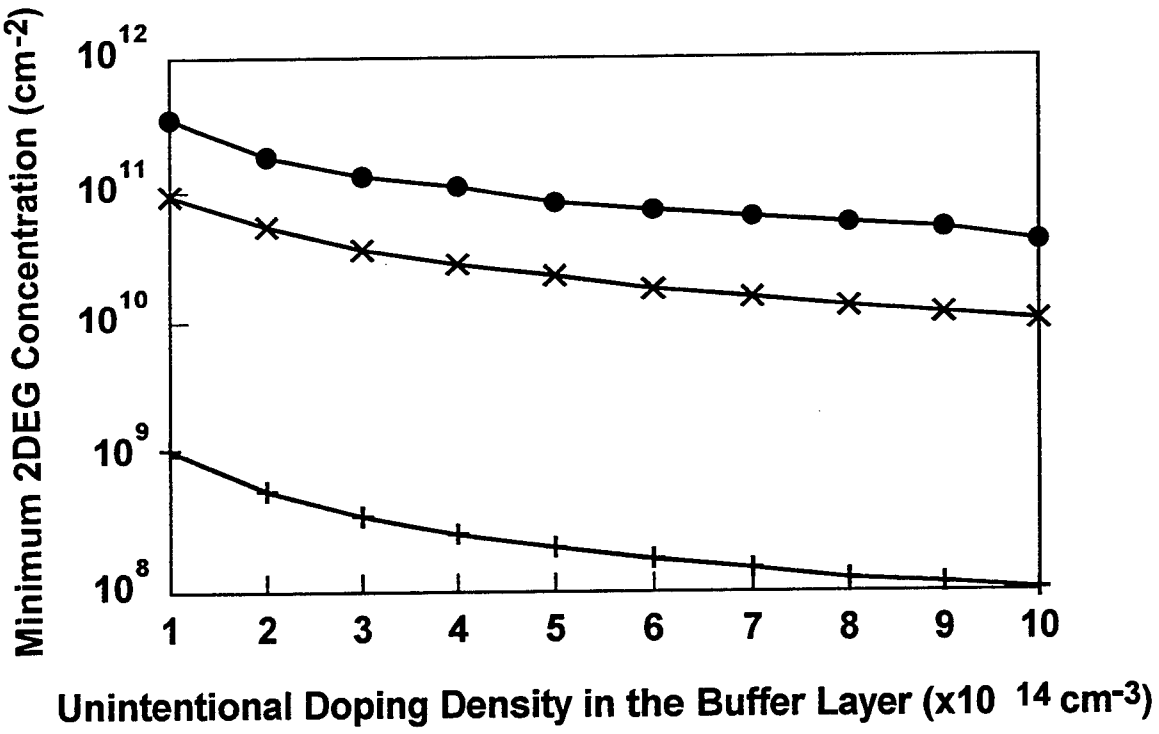


Fig.8 The 2DEG concentration is plotted as a function of the unintentional doping density of the bottom buffer layer at 300K. Well widths of 50A (+), 100A (x) and 150A (•) are shown.

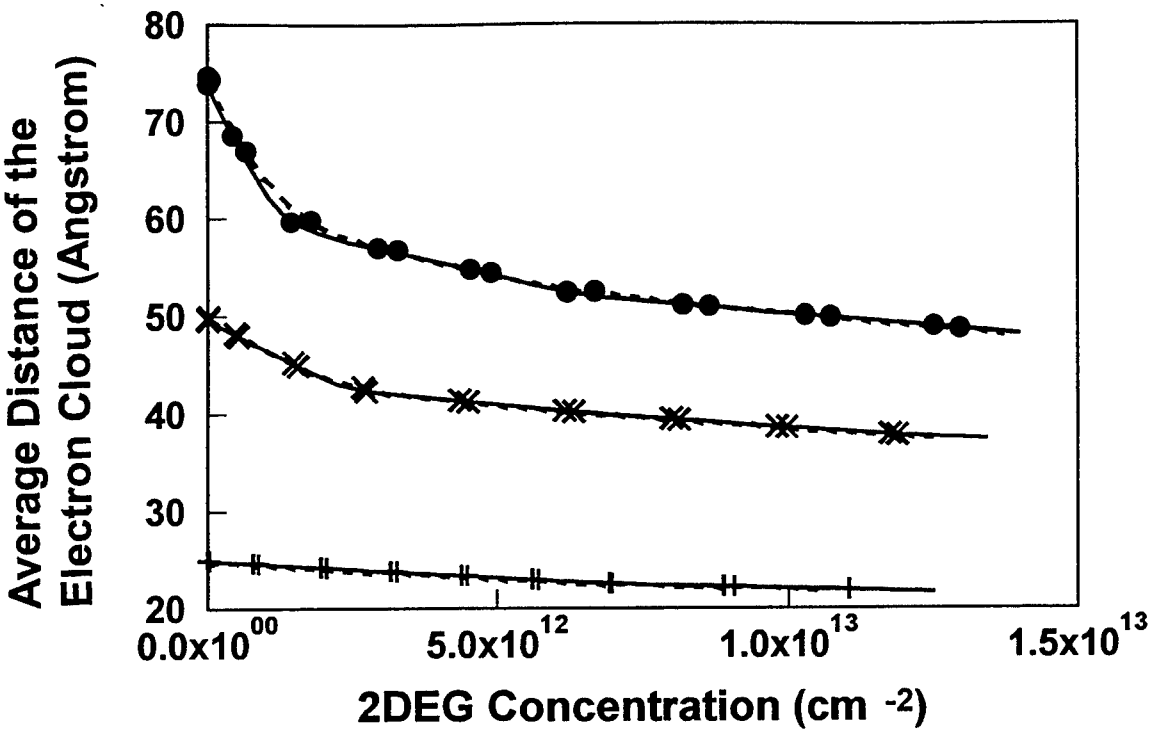


Fig. 9 The average distance of the electron cloud from the first heterointerface X_{av} is plotted as function of n_s for welll widths of 50A (+), 100A (x) and 150A (•), respectively. The solid lines represent 77K and the dashed lines represent 300K.

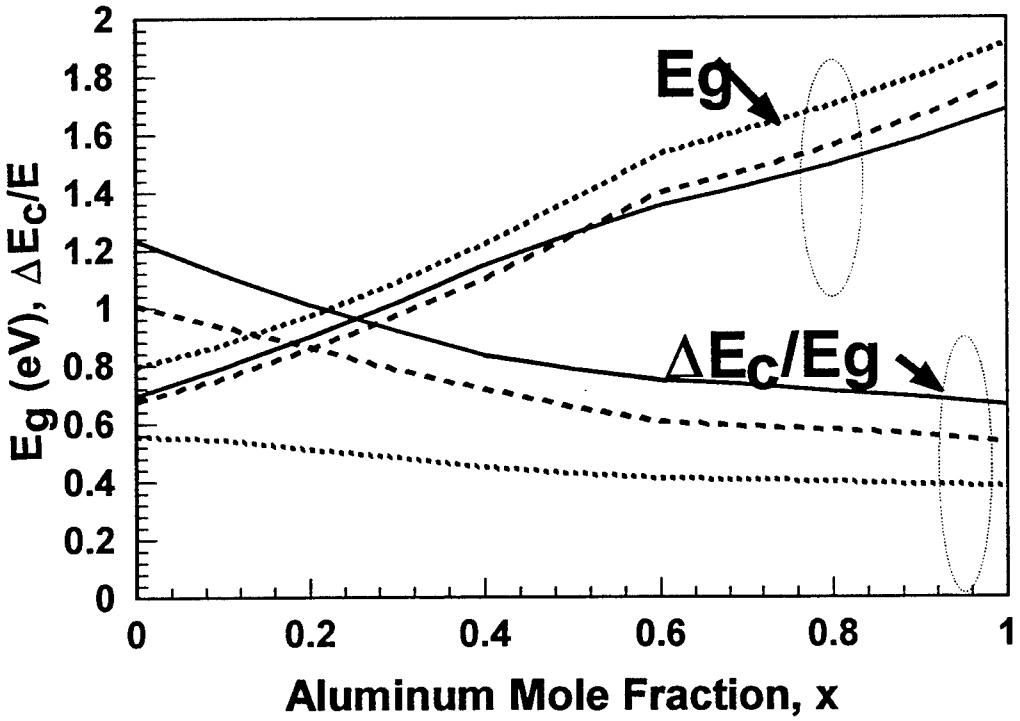


Fig. 10 Energy band gap (solid line) of the quaternary lattice matched to InAs, $In_{.8}Ga_{.2}As$ (long dash) and $In_{.52}Ga_{.48}As$ (short dash).

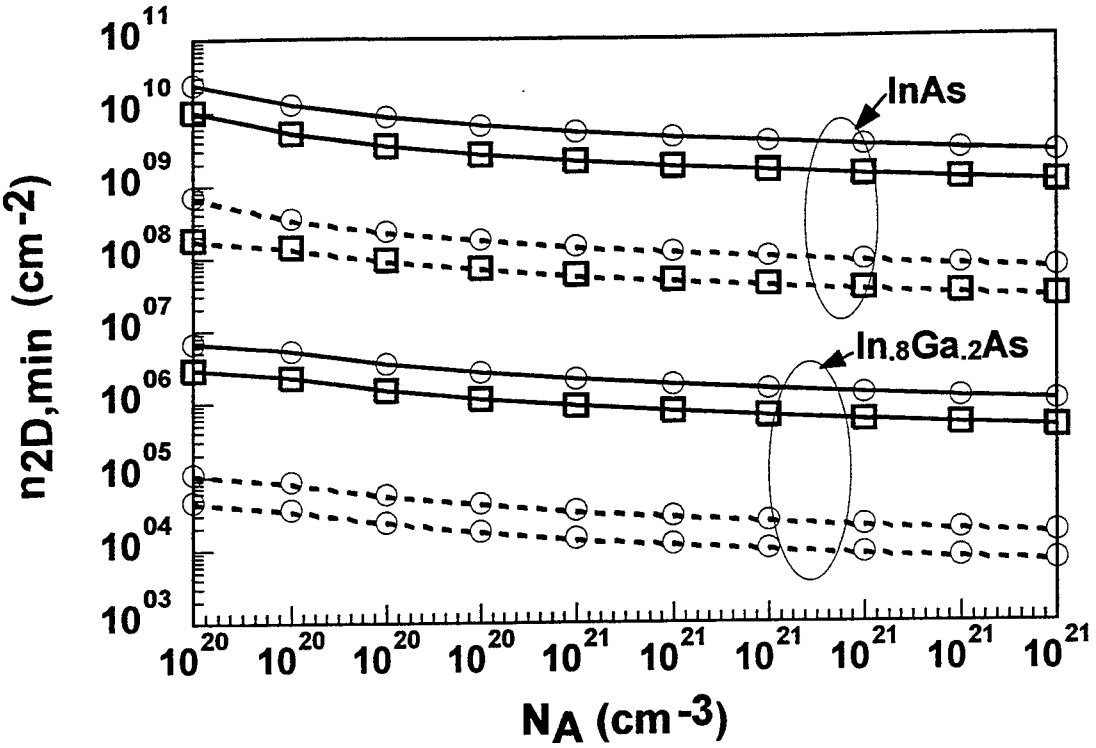


Fig.11 $n_{2D,min}$ vs. N_A with x and the QW width as parameters. Al mole fraction x=0.6 (solid) and x=0.8 (dashed) are considered with QW width of 100A (o) and 150A (□) are considered.

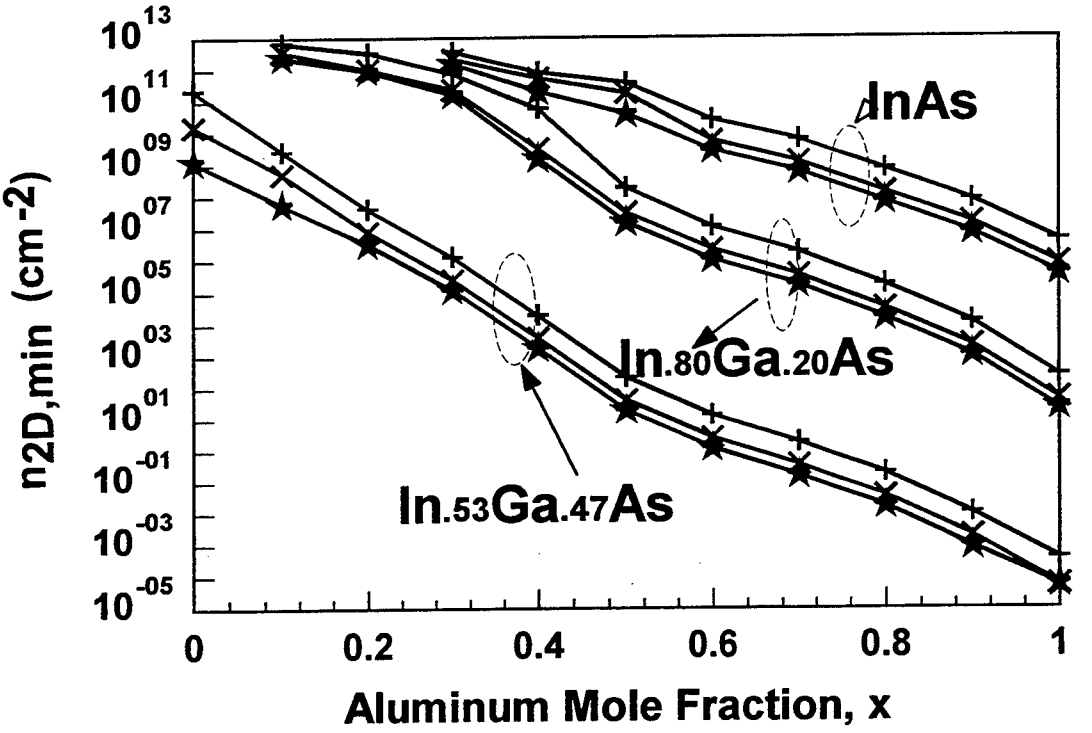


Fig. 12 $n_{2D,min}$ vs. x. Solid line, dashed line and dotted line represent InAs, $In_{.8}Ga_{.2}As$ and $In_{.52}Ga_{.48}As$, respectively. N_A=1x10^{14} (+), 5x10^{14} (x) 1x10^{15} (*) cm^{-3} are considered.

AN ASSESSMENT OF THE CURRENT STATE OF THE ART OF STAP FROM AN ELECTROMAGNETICS POINT OF VIEW

Ercument Arvas
Professor
Department of Electrical Engineering and Computer Science

Syracuse University
121 Link Hall
Syracuse, NY 13244

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific research
Bolling Air Force Base, DC

and

Rome Laboratory

August 1996

AN ASSESSMENT OF THE CURRENT STATE OF THE ART OF STAP FROM AN ELECTROMAGNETICS POINT OF VIEW

Ercument Arvas
Professor
Department of Electrical Engineering and Computer Science
Syracuse University

Abstract

An assessment, from an Electromagnetics point of view, of the state of the art of Space-Time Adaptive Processing (STAP) for radars is given. The validity of certain assumptions made for the antenna system of the radar is discussed. The importance of some of the electromagnetics effects not explicitly included in the STAP algorithms are briefly summarized. These effects include the mutual coupling of the elements of the array, the transmit/receive pattern of individual elements of the array, the near-field-scattering from objects close to the radar, and the effect of a radome housing the antenna. Once the above-mentioned electromagnetic effects are included in a STAP algorithm, one can not assume electrically identical elements of the array antenna. This will further complicate the algorithm.

AN ASSESSMENT OF THE CURRENT STATE OF THE ART OF STAP FROM AN ELECTROMAGNETICS POINT OF VIEW

Ercument Arvas

Introduction

Space-time adaptive processing (STAP) is a multidimensional adaptive filtering algorithm that simultaneously combines the signals from the elements of an array antenna and the multiple pulses of a coherent radar waveform, to suppress interference and provide target detection. Some STAP algorithms developed for airborne radars are reviewed by Ward {1}. Each algorithm is designed based on certain assumptions about the radar system. The antenna is a very important element of any radar. Therefore, an accurate modelling of the antenna system is a crucial part of the overall model of the radar system. Since the electromagnetics of the antennas used in most radars is complicated, usually STAP designers make simplifying assumptions about the antenna of the radar and concentrate on the signal processing part of the system. The purpose of this report is to summarize some of the effects of such assumptions.

Mutual Coupling and the Near-Field-Scattering

Many types of radiating antenna elements can be used in an array antenna of a radar system. However, it is well known that the

properties of an individual element in an array can differ significantly from its properties when it is isolated {2}. In other words, N elements of an array may physically look identical, but it is impossible for them to be electromagnetically identical. For example, the radiation resistance of a thin half-wave dipole radiating in free space is 73 ohms, but in an infinite array backed by a screen it can be 153 ohms. The impedance will also vary with the scan angle. To properly match the elements of the array to the transmitters, receivers and the lines, one must know how the impedance of each element changes.

Therefore, the assumption that each element of the array has identical transmit/receive pattern is not realistic. This is equivalent to assuming that radiating elements of the array are independent of each other. This approximation is not an accurate one in real systems. The current in one element does affect the phase and amplitude of the currents in neighboring elements. Since the currents in each element change with the scan angle, the effect of mutual coupling depends on the scan angle. The dependence of mutual coupling on the change of the beam position in an electronically scanned array results in difficulties in controlling the low sidelobes. This further complicates the problem of computing and compensating for the mutual coupling effects. The diffraction of energy by neighboring elements can also be considered as a mutual coupling effect. This is especially important in end-fire elements closely spaced.

In summary, mutual coupling causes the pattern of the array to differ from that computed assuming independent array elements. For

any STAP algorithm to be realistic, it must properly account for the mutual coupling. With high speed computers and modern numerical computational techniques it is possible to investigate mutual coupling to obtain an accurate pattern of the array.

The diffraction of energy by objects close to the array is another complicated factor that is usually neglected in STAP algorithms. This near-field-scattering effect depends not only where the array is placed but also on the scan angle. The same array placed at the nose of an aircraft will behave differently when it is placed at a side. When the array is placed at the nose of an aircraft, the beam looking ahead will not see the aircraft. As the beam is scanned to a side, the effect of the presence of the aircraft will be noticed. For an array mounted on a side of an aircraft, the radiation pattern will be distorted by the unsymmetrical structure (say cockpit on one side and a wing on the other side) even when the beam is directed in the boresight direction. The diffractions will be more pronounced when the beam is scanned toward a wing.

The final effect of scattering from nearby objects is to change the radiation pattern of the array. An accurate modelling of such an antenna may be obtained by numerical techniques that includes the aircraft as part of the radiating system.

Effect of Radomes

For obvious reasons most of the airborne and ground-based radar antennas are enclosed in a radome. A properly designed radome does not distort the antenna pattern very much.

The effect of a large radome enclosing a ground-based antenna may be predicted by using certain asymptotic techniques. Ground-based radomes are usually built in the shape of a hemisphere. Therefore as the antenna beam is scanned, it is incident on electrically similar parts of the radome. Hence the characteristics of the pattern are not affected by the scan angle. A similar situation exists with rotodomes where the antenna and the radomes rotate together. Rotodomes are used in ground-based as well as in airborne surveillance radars.

The effect of a relatively small radome on an airborne radar antenna is usually not predictable by a simple theory. Usually the antenna beam sees a different portion of the radome as the scan angle is changed. The transmission properties of radome materials varies with the angle of incidence and polarization. Hence the radome can affect the gain, beamwidth, sidelobe level and boresight direction, as well as the VSWR and the antenna noise temperature. The boresight shift can be important in tracking radars.

In summary, the presence of a radome may affect the radiation pattern of the radar antenna. This is especially true for small airborne radomes which are designed not only to meet the electromagnetics constraints, but also to conform the aerodynamic shape of the aircraft. Usually, the electromagnetics is sacrificed to aerodynamics. The STAP algorithms not including the effects of small radomes are not realistic. Unfortunately, the asymptotic techniques, such as ray-tracing, are not applicable to small radomes with corners and edges. On the other hand, even small radomes are too difficult to accurately analyze with other

numerical methods in real time {3}.

Summary

An accurate knowledge of the overall radiation pattern of an array anntenna and the input impedances of the individual elements are perhaps the most important pieces of information needed in designing a STAP algorithm for an airborne radar. An inaccurate knowledge of the pattern will affect the ability of the algorithm to correctly account for the targets, the jammers and the clutter. An accurate knowledge requires a carefull electromagnetic analysis of the overall system: the array, the radome and even the platform. Although some part of the analysis is possible with the modern numerical techniques, the complete system is still too big to be analysed in real time.

The inclusion of the electromagnetic effects in a STAP algorithm will further complicate the algorithm, because the assumption that the array elements are electrically identical will no longer be valid. This would mean that not only the phase, but also the amplitude of target echo signal received will differ from one antenna element to the other. This in turn will make it difficult to define the usual simple "target" steering vector. Similar conclusions can be made about the jammer and clutter steering vectors.

References

1. J. Ward, *Space-Time Adaptive Processing for Airborne Radar,*

Technical Report 1015, MIT Lincoln Lab., Dec. 1994.

2. M. I. Skolnik, *Introduction to Radar Systems,* McGraw-Hill,1980.

3. E. Arvas, U. Pekel, A. Rahhalarabi, and E. Gundogan, "Transmission through a small radome of arbitrary shape," IEE Proc., Pt. H, vol.137, No.6, pp.401-405, Dec.1990.

Formal Verification Using ORA Larch/VHDL Theorem Prover

Ahmed E. Barbour
Associate Professor
Department of Mathematics and Computer Science

Georgia Southern University
Statesboro, GA 30460-8093

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC

and

Rome Laboratory

October 1996

Formal Verification
Using
ORA Larch/VHDL Theorem Prover

Ahmed E. Barbour

Associate Professor
Mathematics and Computer Science Department
Georgia Southern University

Abstract

This report describes the research conducted in the field of formal hardware verification. Odyssey Research Associate Inc. (ORA) has develpoped a hardware verification environment Larch/VHDL that includes an interactive theorem prover (Penelope). Larch/VHDL is used to prove that certain classes of logic circuits can be verified in a systematic way. Accordingly, a step-by-step verification methodology is being developed to describe the process of performing a formal proof that a VHDL design at the logical and components level satisfies its specification. An example of a full adder designed using AND, OR, NOT, and XOR logic gates has been developed and proved to meet its design specifications using the developed methodology. The same proof steps are applied to prove the correctness of more complex logic structures (a four-bit adder with overflow). The same process can be used to generate algorithms for automating the proof process. Also, proposed research projects based upon the Larch/VHDL tool to make it more acceptable and usable by both universities and industry are being investigated. These projects include developing automated proof at the gate level, establishing fault models and developing its formal specifications, developing formal specifications for the testability of a logic circuit, generating test patterns, and parallelizing Larch/VHDL to reduce its execution time.

Formal Verification
Using
ORA Larch/VHDL Theorem Prover

Ahmed E. Barbour

1. Introduction

During the 1990's, computationally demanding applications have depended on complex architectures to deliver the required performance. Products have increased in functional complexity, with fault-tolerant, and higher overall performance. On the other hand, minimizing the size, weight and power of embeded computer systems depend upon our understanding of the application problem and its computing requirements. This requires analysis, simulation and prototyping of hardware and software. Analysis and prototyping are becoming more expensive and time consuming with increasing design complexity and limited budget. Hence, simulation and formal verification are the only resonable alternative tools which can be used effectively to verify complex systems.

VHSIC Hardware Description Language (VHDL) which was established as an IEEE standard for the design and documentation of digital electronic systems. VHDL was developed in response to the Computer Aided Design (CAD) community's need to handle larger and more complex designs and the need to be able to electronically exchange design information [1-4]. Two methods exist for the verification of a logic circuit's correctness: simulation and formal verification. Simulation techniques use exhaustive testing of the VHDL model on several levels to determine the correctness of a design. The second approach, formal verification of hardware, is a mathematical proof that the design of a digital circuit which satisfies certain properties regardless of the values of the inputs. Figure 1 shows the two approaches to hardware design verification. Formal verification tools can also be used to prove that hardware designs satisfy properties such as functional correctness, security, and timing correctness.

The report is organized as follows. Brief description of both simulation techniques and formal verification are given in Sections 2 and 3, respectively. Section 4 explains the ORA Larch/VHDL hardware verification environment (Penelope Theorem Prover) and the methodology used by Larch/VHDL to construct formal proof. A systematic and step-by-step method to construct formal proof for a four bit adder with overflow using four full

adders; each constructed by logical gates: AND, OR, NOT, and XOR is developed. The examples show the possibility to automate the proof for a complex gate level structure. Section 5 briefly describs research projects which could be conducted in the future to improve the performance of the theorem prover. Section 6 concludes the most important issues addressed in this report.

2. Verification By Simulation

VHDL is the IEEE standard (IEEE Std 1076-1993) representation for modeling and simulating digital circuits [1-4]. The use of VHDL in a top-down design provides many benefits including modeling at multiple levels of abstraction, technology indpendence and validation through simulation [15]. Figure 2 shows the basic components for VHDL verification using simulation approach. WAVES (IEEE Std 1029 and Std 1164-1993, Multivalued Logic System) is the industry standard representation and exchange format for digital stimulus and response data [2]. The words "waveform" and "vector" indicate that WAVES may represent simulator event trace data, as well as the highly structured test vectors typical of automated test equipment. The word "exchange" means that WAVES is meant for the exchange of information between vendors as well as design and test environment [16]. WAVES is a subset of the IEEE Std 1076-1993, also known as VHDL.

VHDL was chosen as the basis for WAVES because VHDL is so important in the design phase of electronic components and modules. WAVES provides a powerful support mechanism for concurrent engineering practices by allowing digital stimulus and response information to be freely exchanged between multiple simulation and test platforms. WAVES is defined as a syntactic subset of VHDL, and as much, can be simulated against the VHDL model during design to verify the functionality and timing of the design as it progresses. Further, when devices are fabricated, the same WAVES data set can be used in the electrical test process to assure that the same stimulus that was used during design is applied after the fabrication process and during electrical test [15,16]. Figure 3 shows how to use WAVES as a test bench to verify a VHDL design.

3. Formal Verification Using Larch/VHDL

The US Air Force's Rome Laboratory contracted ORA to develop a formal hardware verification environment. The goal of this contract is to develop the capability to logically prove that a VHDL hardware design is functionally correct over all possible input combinations. Likewise, it is essential that there exists an ability to prove the absence of certain properties, e.g. deadlock, or resource contention. ORA is leveraging off an Ada

verification environment they have developed, known as Penelope [5-8]. Penelope is based upon the Larch two-tiered specification language developed at the Massachusetts Institute of Technology (MIT) [9-12]. The first tier, the Larch Shared Language (LSL), is a first order predicate calculus used to build the traits, or theories, that define the sorts (or in the case of VHDL, types) used by the target language, i.e., bit, word, string, arrays, integer, etc. The second tier, called the Interface Language, defines the communication mechanisms of the target language, Ada, C, C++ or in this case VHDL, in the Larch notation. LSL is used to mathematically model data objects and operations on those objects, while the interface language maps the VHDL model into the abstractions represented by the Larch expressions for the purpose of formal reasoning. Figure 4 shows the structure of Larch/VHDL formal proof methodology.

Larch/VHDL verification environment is an interactive tool that helps its user to develop and verify digital electronic hardware designs written in VHDL. Larch/VHDL tool is well suited to developing code in the goal-directed style advocated by Gries [13] and Dijkstra [14]. In this style the designer develops a VHDL model from a specification in a way that ensures the VHDL model will meet the specification. Larch/VHDL supports a window interface environment with several advanced features for entering specifications, developing code, and providing access to the Penelope theorem prover. Figure 5 illustrates the internal structure of Penelope system. Some features of the Larch/VHDL design verification process are [6-8]:

(1) Larch specifications are independent of technology and implementation details. The characteristic of these specifications are: (a) unambiguous, concise and immune to errors, (b) proven to be correct, (c) combined to form new specifications, and/or (d) implemented in hardware or software.

(2) Formal verification increases the designer's confidence.

(3) Multiple implementations of a design in VHDL are possible to be verified. In many cases, several VHDL architectures may be developed in order to perform tradeoff analysis. All these VHDL models will conform to the same entity interface declaration and its specification.

(4) Hierarchical Verification is supported. The Larch/VHDL methodology supports a form of verification of single components or cells of a design which is done earlier in the development cycle than simulation is typically done. Once verified, these components are available for reuse in other designs.

4. Larch/VHDL Formal Verification Environment

The Larch/VHDL environment includes a large body of traits that define the basic constructs of digital design such as bit, vector, gate, logic operations and so on. Traits define sorts (logical types) and state properties or assertions that must hold true. Traits also contain theorems which are statements that are deducible from assertions, previously deduced theorems, and/or the assertions or theorems of other traits that are included. The two-tiered Larch approach allows designers the capability to extend the library of traits in order to support user defined sorts in their models. Once implemented, the traits are available as library components for reuse in other applications. Traits are used to capture the concepts and relationships used in digital design. There are traits devoted to arithmetic concepts, and to data structures such as arrays and lists. To support VHDL semantics there are traits defining signals, and signal delay, and other concepts needed to express the semantics of VHDL. There are traits that describe the relationship between bit level operations and their arithmetic interpretation, in twos-complement or unsigned bit-level representations.

4.1 Penelope's Proofs Methodology

Penelope Larch/VHDL theorem prover includes a simple proof editor/checker for predicate calculus that provides a number of proof rules for performing simplification and proofs. Penelope applies the rules according to user directions and indicates to the user what, if anything, still has to be proved after each step. Each statement to be proved or simplified is presented in the form of a sequent, a set of hypotheses and a conclusion. Each proof in Penelope takes place in the context of an available theory. Within a VHDL design unit, the theory is determined by entity declaration annotations and all the local lemmas currently being applied to complete the proof. The theory that is available for proving a given lemma consists of the axioms, assumptions, and proved lemmas that precede the given lemma. Figure 6 shows the general relationship among entity, architecture, specifications, and theories used to help the proof construct. One entity could have many VHDL forms (architecture, behavior, or logical); each form has its own proof section; each proof could use new theories to support the proof process. The specification section of the proof could also use new theory [6-8].

Penelope's proofs are tree-structured. Each node of the tree corresponds to a sequent to be proved and one proof step that proves it, possibly subject to proving other, derivative sequents. The children or subproofs of the node correspond to the further sequents needed to prove it. Leaves of a completed proof correspond to

sequents that require no further proof (e.g., a sequent whose conclusion is "true"). While constructing a proof, the leaves also include unproved sequents. The proof rules are organized with a hierarchical menu. When the cursor is positioned at a proof step, each item on the menu may represent a proof rule or a group of proof rules (for example, thinning is a proof rule, but analyze-hypothesis represents a group of proof rules). If the help-pane menu item corresponds to a single proof rule, clicking on the menu item causes the proof rule to be added to the proof tree. If a group of proof rules is chosen, a submenu appears with the individual rules in the group. The large number of proof rules available may make the Penelope prover seem formidable. Penelope's proof steps fall into several basic groups: the application of automatic simplifiers or rewriters; the application of some available theorem (called instantiation); rules (such as and-synthesis, mentioned above) that break down the conclusion or hypotheses according to their syntactical form; and proof-structuring rules, such as proof by cases or proof by induction [6-8].

4.2 Verification Examples

Large number of VHDL design verification examples have been established to illustrate the structure of the formal proof (see Ref. [17-19] and my home page: http://gsu.cs.gasou.edu/~barbour). Four examples have been selected to illustarte the decomposability property of the proof, to show the effect of logic circuit with delay, and the problem of proving iterative logic circuits. The first example is a full adder which is constructed from AND, OR, NOT and XOR logic gates which have been proven to be correct. The second example is a four-bit-ripple adder which consists of four full adders with an overflow signal as shown in Figure 7. A step-by-step proof explanation is given to illustrate the basic element of the proof process. It is very interesting to notice that the same proof methodology used to proof a full adder is also used to verify four-bit adder using four full adders. The examples are not complicated VHDL design but the information provided by the definition of VHDL semantics can be shown and the verification conditions that must be satisfied to meet the specification can be seen. In each example, the designer always provides three declarations: (1) an entity declaration, (2) an architecture for the entity, and (3) a specification written in Larch/VHDL as shown in Figure 6.

Example 1: A Full Adder

VHDL Entity Declaration

```
entity full_adder is port(x, y, ci :bit; s, co : out bit); end full_adder;
```

VHDL Architecture Description

```
architecture arch of full_adder is
component and2  port(x,y : Bit; z : out Bit); end component;
```

```
component ex_or port(x,y : Bit; z : out Bit); end component;
component or2   port(x,y : Bit; z : out Bit); end component;
signal  z1, z2, z3 :bit;
begin
l1: ex_or port map(x,y,z1);        l2: ex_or port map(ci,z1,s);  l3: and2 port map(x,y,z2);
l4: and2 port map(z1,ci,z3);       l5: or2 port map(z2,z3,co);
end arch;
```

Larch/VHDL Specification

```
entity full_adder
guarantees      always  s =  ( x xor y xor ci) with s  delayed by 0 from x,y,ci
guarantees      always  co = ( ( (x and y ) or (ci and (x xor y))))
                        with co  delayed by 0 from x,y,ci
end
```

Since we are interested in the proof methodology, the Verification Conditions (VCs) generated by the Larch/VHDL tool are removed due its length and can be found in my home page. It is important to notice that each logical gate (ex_or, and2, or2) has its own proof structure which is used by the proof structure of the full adder. After inserting obligation, the step by step proof is shown below.

Larch/VHDL Proof

```
/* Library used by Larch/VHDL Prover  */
--| library lib --|        WORK --|        STD --|        INITIAL_THEORY
--|        VHDL_MATH --|        GENERAL_MATH ;
--| proof:
Step (1): Insert Obligation which inserts Verification Conditions (VCs) generated
          by the prover.  These VCs should be proven to be true by the designer.
Step(2): Use "Synthesis-Conclusion" and Select "Forall/Implies" as shown.
            BY synthesis of FORALL/IMPLIES
Step (3): Use "Sythesis_Conclusion" and Select "And-Synthesis", the prover will split
          the proof into two part as shown.
            BY synthesis of AND
Step (4): For the first part, Use "Synthesis-Conclusion" and Select "Forall/Implies" and
          then Select "with Analysis", the prover will simplify the first part of the proof.
             BY synthesis/analysis of FORALL/IMPLIES
Step (5): Use "Thinning" and Select "Binding" and then type: "*", the prover will
          simplify it more.
             BY thinning (binding *)
Step (6): Use "Simplify" and Select "SDVS Simplify" and then Select ""+".
          the prover will complete the proof of part 1.
             BY simplification+
             BY synthesis of TRUE
Step (7): For the second part, Select "Synthesis-Conclusion" and Select "Forall/Implies"
          then Select "with- Analysis", as shown.
             BY synthesis/analysis of FORALL/IMPLIES
Step (8): Select "Thenning" and Select "Binding" then type: "*", the prover will
          simplify the proof.
             BY thinning (binding *)
Step (9): Select "Simplify" and Select "SDVS-Simplify" then Select "+", and
          Select "Rewriting" and then Select "+", and finally Select "SDVS-Simplify",
          the prover will complete the proof.
             BY simplification+, rewriting+, simplification
             BY synthesis of TRUE
```

Example 2: A Four-bit Ripple Adder with Overflow

A four-bit ripple adder is constructed using four full adders. The overflow signal is drived from the XOR function of the two last carries as shown in Figure 7.

VHDL Entity Declaration

```
entity adder4_ovf is port(x1, y1, x2, y2, x3, y3, x4, y4, ci :in Bit;s1, s2, s3, s4, c1, c2, c3, co : out Bit);
end adder4_ovf;
```

VHDL Architecture Declaration

```
architecture arch of adder4_ovf is
component full_adder     port(x,y,ci : Bit; s,co : out Bit); end component;
component ex_or          port(x,y : Bit; z : out Bit); end component;
begin
I1: full_adder port map(x1,y1,ci,s1,c1);  I2: full_adder port map(x2,y2,c1,s2,c2);
I3: full_adder port map(x3,y3,c2,s3,c3);  I4: full_adder port map(x4,y4,c3,s4,co);
I5: ex_or port map(c3,co,ovf);
end arch;
```

Larch/VHDL Specification

```
entity adder4_ovf
guarantees  always s1 =  ( x1 xor y1 xor ci) with s1  delayed by 0 from x1,y1,ci
guarantees  always c1 = ( ( (x1 and y1 ) or (ci and (x1 xor y1)))) with c1  delayed by 0 from x1,y1,ci
guarantees  always s2 =  ( x2 xor y2 xor c1) with s2  delayed by 0 from x2,y2,c1
guarantees  always c2 = ( ( (x2 and y2 ) or (c1 and (x2 xor y2)))) with c2  delayed by 0 from x2,y2,c1
guarantees  always s3 =  ( x3 xor y3 xor c2) with s3  delayed by 0 from x3,y3,c2
guarantees  always c3 = ( ( (x3 and y3 ) or (c2 and (x3 xor y3)))) with c3  delayed by 0 from x3,y3,c2
guarantees  always s4 =  ( x4 xor y4 xor c3) with s4  delayed by 0 from x4,y4,c3
guarantees  always co = ( ( (x4 and y4 ) or (c3 and (x4 xor y4)))) with co  delayed by 0 from x4,y4,c3
guarantees  always ovf = (c3 xor co) with ovf delayed by 0 from c3, co
end
```

It is clear from the above specification that statement "guarantees .." repeats itself. A sample of step by step proof is shown below which indicates the simplicity of the proof due to the iterative nature of the specification.

Larch/VHDL Proof

BY synthesis of FORALL/IMPLIES **BY synthesis of AND**
BY hypothesis BY hypothesis BY hypothesis BY hypothesis BY hypothesis
BY hypothesis BY hypothesis BY hypothesis BY hypothesis

It is important to notice that only two lines of proof (bold face statements) are needed to proof this complex structure. Therefore, the great potential for theorem proving to verify a complex hierarchical design is available to the user as shown in the previous two examples. The following example, Example 3, illustrates the effect of delay introduced by signals propagated through the logic gates. The delay is represented by a positive integer value called: **del** which represents the delay of a signal logic gate: AND, OR, NOT, or XOR. Example 3 shows the carry signal of the full adder discussed in Example 1 with a del applied to each gate.

Example 3: Carry Logic Circuit with Delay

Entity Declaration

```
entity carry_del is generic (del : positive); port(x, y, ci :bit; co : out bit);
end carry_del;
```

Architecture Declaration

```
architecture arch of carry_del is
```

```
component and2_del generic (del : positive); port(x,y : Bit; z : out Bit); end component;
component or2_del generic (del : positive); port(x,y : Bit; z : out Bit); end component;
signal  z1, z2, z3 :bit;
begin
  11: or2_del generic map(del) port map(x,y,z1);  12: and2_del generic map(del) port map(x,y,z2);
  13: and2_del generic map(del) port map(z1,ci,z3); 14: or2_del generic map(del) port map(z2,z3,co);
end arch;
```

Specification Declaration

```
entity carry_del
guarantees  always co = ( ( (x and y.) or (ci and (x or y)))) with co  delayed by 3*del from x,y,ci
end
```

The proof is similar to the proof of Example 1 with only one statement to be added: **del_constraint,** as shown.

Proof Section

```
BY synthesis/analysis of FORALL/IMPLIES  BY thinning (binding *)
BY using del_constraint as new hypothesis  BY simplification+  BY synthesis of TRUE.
```

As shown in the previous examples, the proof can be constructed to become more systematic for most logic circuits. However, changing the architecture of four-bit adder to become iterative using Bit_vector to represent its input and output signals and using for statement to represent the iterative nature of the full adder components, and changing the specification to be more compact, the proof becomes more complex and difficult to be achieved in a systematic way. Example 4 shows the iterative form of the architecture and the specifications of a four-bit adder using for loop to repeat the structure of the four full adders.

Example 4: Iterative Four-Bit Ripple Adder Using For Loop

Entity Declaration

```
entity add_4 is    port ( x, y : Bit_vector (3 downto 0); ci : Bit_vector (4 downto 0);
                          s, co : out Bit_vector (3 downto 0));        end add_4;
```

Architecture Declaration

```
architecture arch of add_4 is
component full_adder port(x,y,ci : Bit; s,co : out Bit); end component;
begin     d0to3 : for i in 0 to 3 generate
            d0   : full_adder port map(x(i),y(i),ci(i),s(i),co(i));
            ci(i+1) <= co(i);
        end generate;
end arch;
```

Specification Declaration

```
entity add_4
includes ( BitVector)
guarantees  always forall i : Int :: 0 <= i and i <= 3 ->
                  s[i] =  (x[i] xor y[i] xor ci[i]) with s delayed by 0 from x, y, ci
guarantees  always forall i : Int :: 0 <= i and i <= 3 ->
                  co[i] = ( ( (x[i] and y[i]) or (ci[i] and (x[i] xor y[i])))))
                  with co delayed by 0 from x, y, ci
guarantees  always forall i : Int :: 0 <= i and i <= 3 ->
                  ci[i+1] = co[i]     with ci delayed by 0 from co
end
```

The above iterative four-bit adder structure has the same logical structure of Example 2; the only difference is using for loop to repeat the full adder. Large number of examples have been constructed and proved using

high level behavior specifications of logical system. However, few examples are given using low level gate structure of the system. The importance of the gate level representation of a logical system becomes more critical when modeling faults (stack-at or bridging faults), generatting test pattern, or verifying the testability of a logic circuit. These critical formal verification issues are discussed in the next section.

5. Suggested Research Projects

The complexity of the formal verification for even resonable size of hardware design makes it difficult to be used on wide scale in industry and universities. From my experience in using Larch/VHDL theorem prover during the Summer Research Program, the following research projects have been recognized as important research issues to be investigated in the future to improve the performance of the theorem prover, to make the tool more efficient and applicable to a wide range of logical systems.

5.1 Automated Proof

The main drawback of Larch/VHDL theorem prover is its limited capability to automate the proof. The examples given in Section 4 illustrate the capability inherent in the prover to proof complex structure using the proof of its components. The example shown in Figure 8 indicates the possibility to automate the proof of of an XOR logic function. The proof may break down to its simple elements of repeated application the assignment statement: z <= x and the logical operators AND, OR, and NOT. So if the proof of the assignment statement and the proof of each logical operation are constructed, then the formal verification of the XOR function follows the tree structure shown in Figure 8. Also, it is known that the theorem prover follows a tree structure form in its proof which can be easily automated.

5.2 Test Bench Generator By Theorem Prover

The final process of any logical design is to implement the design and produce the required hardware device (IC or Microprocessor). One part of the overall system quality assurance process is the testing of the manufactured system for design defects. A test design strategy to reveal faulty systems at the time of manufacture is required to assist the designer in demonating overall system quality [23]. Then the most important question is: how to test a logical system after proving and verifying its correct behavior that it statisfies its design specificaton? Even with successful formal proof of the logic circuits, it is imperative to test the device using test patterns. The possibility to make a theorem prover generates test patterns for the device under verification

is inherent in its process. To be able to do so, it is required to establish the following specifications: (a) fault model inside the VHDL description, (b) the capability to inject faults into specific part of the logic circuit, (c) generate specification conditions, (d) verify the testability of a VHDL model, and (e) finding algorithm to generate test patterns while searching for a proof.

Fault simulation techniques which are applicable to VHDL models are still in its early stage [24,25]. A fault simulator is used to determine which faults are detected for a given set of test patterns. In this regard, the report submitted by B. W. Jhonson, D. Todd Smith, and T. A. DeLong to Rome Laboratory/PKRZ establishes the basis for advance research in the field of fault simulation, VHDL, testing and testability of logical systems [23]. A very simple example which illustrates the effect of static hazard on the performance of a logic circuit is shown in Figure 9. In this simple circuit, it is important to proof that the circuit produces hazard and the hazard free design is testable [20,21].

5.3 Formal Verification of Sequential Systems

Most of today logical systems contain sequential circuits in their logical structure. Therefore, formal verification of a sequential machine using its state table or diagram is a very important requirement of any theorem prover. As an example, Figure 10 shows a simple finite state machine which detects the sequence 0101 in an input string of 0s and 1s. Penelope theorem prover is unable to proof such system [20]. Therefore, it is important to address this issue and be able to proof the correctness of the state diagram or the state table of the suggested machine.

5.4 Formal Verification of Fault-Tolerant Systems

Some work has been done before by ORA to verify FtCayuga Fault-Tolerant Microprocessor System [22] using different tools and different methodology. The importance of fault-tolerant systems is increasing every day due to the increased complexity of hardware and software. The verification that a fault-tolerant system satisfies its specification and it is testable at the same time is a very important design issue in this field. A fault-tolerant system could be any one or combinaton of the following designs:

system level, gate level, parallel system, error detection, and error correction system.

5.5 Parallel Implementation of Theorem Prover

Due to the complexity of the formal proof, it is possible for a certain proof to takes hours or days. The parallel implementation of the theorem prover becomes more important in today technology. Larch/VHDL theorem prover lends itself to parallel implementation on the verification conditions level and on the proof level as shown in Figure 6. Using PVM (Parallel Virtiual Machine) software package developed at Oak Ridge National Laboratory, Oak Ridge, TN, several Sun workstations connected by Ethernet Network can be used as a message passing parallel system [26]. Figure 11 shows the structure of PVM and the structure of the formal proof which indicates the possibility of implemented the proof in parallel using PVM and network of Sun worksations.

5.6 Education in Formal Verification

The acceptance of a standard, formal specification and verification as a common practice is slow and required long process involving user demand, support tooling, user acceptability, and more important, user understanding and appreciation to its benefits. The main ingredient to the adoption of a standard as a common practice is the availability of information for the use and application of the standard. I am very concerned of how formal verification can be introduced to small universities and colleges. In this part of the research, it is important to make formal verification easy to be used and more understanble to a wide range of engineers and scientists, not only to a limited number of experts in this field. The only way to achieve this objective is to make the theorem prover as a tool accecable to all universities and industry (technology transfer), to establish a methodology for building a proof, and highlighting the importance of the formal specification and verification methodology by offering regular workshops, technical sessions, tutorial conferences, and learning support facility on a wide scale.

6. Conclusions

Larch/VHDL is an interactive environment that helps its user to develop and verify digital electronic hardware designs written in VHDL. The Larch/VHDL environment provides a means to specify and verify a hardware design prior to simulation and in a manner that supports specification and design reuse. Larch/VHDL can also be used to verify previously written VHDL models by developing the Larch specification for the models. Larch/VHDL is the user's trained assistant in verification. It performs well-defined but tedious tasks, like computing verification conditions and carrying out proof steps, while the user is responsible for the intelli-

gent part of the work, specifying the design, developing the design, and deciding how to prove it. Unfortunately, all but the most trivial simplification and proof in Penelope require the guidance and control of the user. This interaction is necessary because of the well-known fact that simplification and theorem proving are in general undecidable.

During my Summer research period, large number of logic circuit design examples were established and verified step-by-step using Larch/VHDL theorem prover. Each proof was established with delay and without delay. The difficulty of proving iterative logic circuits using for statement was reported and systematic approach to prove similar problem will be established in the future. Several research issues relating to improve the performance of the theorem prover have been suggested. The most important issue is how to make formal verification more understandable and easy to use tool and how to introduce it into small universities and colleges.

7. Acknowledgement

I would like to thank all the stuff members of AFOSR Summer Research Program Office and the Rome Laboratory/ERDD stuff members for their help and assistant during my research period. Specific thanks are due to R. G. Hillman, ERDD Branch Head, for his support and encourgement, to Dr. E. Stabler for his constant effort to make theorem prover more applicable, and to M. Nassif, my focal point, due to his valuable time and effort which made my research to be productive and my life easy and enjoyable.

8. References

[1] IEEE Standard VHDL Language Reference Manual. ANSI/IEEE Std 1076-1993. June 4, 1994.

[2] IEEE Standard for Waveform and Vector Exchange (WAVES). IEEE Std 1029.1-1992, and IEEE Standard Multivalue Logic System for VHDL Interoperability (std_logic_1164), May 26, 1993.

[3] Z. Navabi, VHDL: Analysis and Modeling of Digital Systems, McGraw-Hill, 1993.

[4] P. J. Ashenden, The Designer's Guide to VHDL, Morgan Kaufmann, 1996.

[5] D. Jamsek and M. Bickford, "Formal Verification of VHDL Models," Final Technical Report, Rome Laboratory RL-TR-94-3, March 1994.

[6] Odyssey Research Associates, "Penelope Reference Manual V3-3," TM94-0009, December 1993.

[7] M. Bickford, "Technical Information Report: Final Report for Formal Verification of VHDL Desin," Odyssey Research Assocates, F30602-94-C-0136, CDRL, A005, Rome Laboratory/ERDD, July 1996.

[8] M. Bickford, "Technical Information Report: User/Training Manual for Formal Verification of VHDL Design," Odyssey Research Assocates, F30602-94-C-0136, CDRL, A004, Rome Laboratory/ERDD, 621 July 1996.

[9] S. Garland, J. Guttag and J. Horning, "Debugging Larch Shared Language Specification," IEEE Trans. on Software Engineering, Vol. 16, no. 9, September 1990.

[10] J. Guttag, J. Horning and J. Wing, "The Larch Family of Specification Languages," IEEE Software, September 1985.

[11] J. Wing, "Writing Larch Interface Language Specifications," ACM Trans. on Programming Languages and Systems, Vol. 9, no.1, January 1987.

[12] J. V. Guttage and J. J. Horning, LARCH: Languages and Tools for Formal Specification, Spring-Verlag, 1993.

[13] David Gries, The Science of Programming. Springer-Verlag, 1981.

[14] E. W. Dijkstra, A Discipline of Programming. Prentice Hall, Englewood Cliffs, 1976.

[15] C.J. Flynn, F. G. Hall, J.P. Hanna, M.P. Nassif, and M.T. Pronobis, "Top-Down Design Trough Test Using VHDL and WAVES, " Proc. of the 1995 Western MultiConference, Int. Conf. on Elect. Hardware Description Languages (ICHDL), Las Vegas, NV, pp. 21-25, Jan. 15-18, 1995.

[16] S. Drager, C. Flynn, F. Hall, J. Hanna, R. Hillman, and J. Nagy, "WAVES-VHD Integration for Common Applications, "In-House Report RL-TR-96-3, Rome Laboratory, Air Force Material Command, Rome, NY, Feb. 1996.

[17] M. P. Nassif and R. J. Paragi, "The Larch/VHDL Methodology for Hardware Verification," Proce. of the Second ISSAT Int. Conf. on Reliability and Quality in Design, Orlando, Florida, pp. 243-247, March 8-10, 1995.

[18] E. P. Stabler, M. P. Nassif, and R. J. Paragi, "Extending The Design Process With Formal Verification Technology," to be appeared at the Proc. of the VHDL Int. Users Forum (VIUF), Santa Clara, CA, Feb. 28- March 2, 1996.

[19] E. P. Stabler, M. P. Nassif, and R. J. Paragi, "Verification of ASIC Designs in VHDL Using Computer-Aided Reasoning," to be appeared at the Proc. of the Sixth Annual Dual-Use Technologies and Applications Conf., Syracuse, NY, June 3-6, 1996.

[20] Z. Kohavi, Switching and Finite Automata Theory, 2nd. Ed., McGraw-Hill, 1978.

[21] E.J. McCluskey, Logic Design Principles with Emphasis on Testable Semicustom Circuits, Prentice Hall, 1986.

[22] M. Srivas and M. Bickford, "Verification of the FtCayuga Fault-Tolerant Microprocessor System," Vol. 1, A Case Study in Theorem Prover-Based Verification, ORA Corporation, NASA Contractor Report 4381, July 1991.

[23] B.W. Johnson, D.T. Smith, and T.A. DeLong, "A Survey of Fault Simulation, Fault Grading, and Test Pattern Generation Techniques with an Emphasis on the Feasibility of VHDL Based Fault Simulation," Report submitted to Rome Laboratory by the Center of Semicustome Integrated Systems, University of Virginia, Charlottesville, Virginia, May 10, 1996.

[24] V. Pitchumani, P. Mayor, and N. Radia, "A System for Fault Diagnosis and Simulation of VHDL Descriptions," Proc. of the 28th ACM/IEEE Design Automation Conference, San Francisco, CA, pp. 145-150, June 17-21, 1991.

[25] Z. Navabi, N. Cooray, and R. Liyanage, "Modeling for Fault Insertion and Parallel Fault Simulation," Proc. of the VIUF, Spring 1993, pp. 75-89, April 28-30, Chateau Laurier, Ottawa, Canada, 1993.

[26] A. Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Mancheck, and V. Sunderam, PVM 3.0 (Parallel Virtual Machine) User's Guide and Refrence Manual, Oak Ridge National Laboratory, Oak Ridge, TN, Feb. 1993.

Figure 1: Hardware Design Verification Techniques

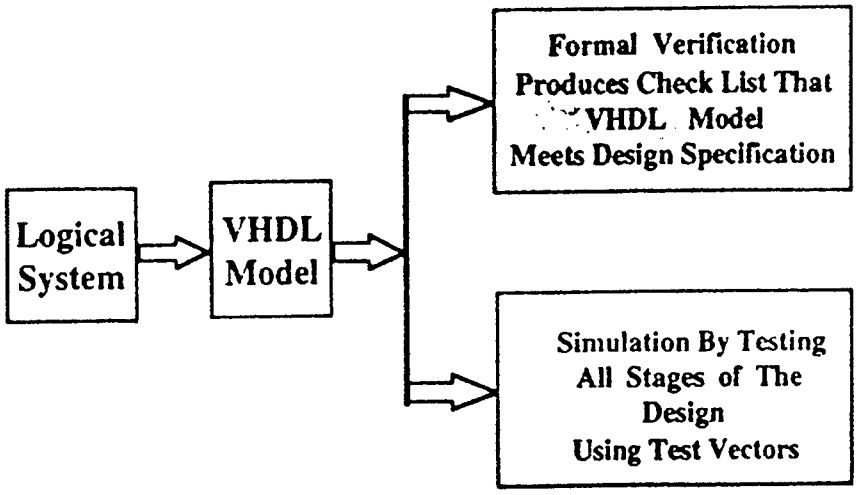


Figure 2: Verification by Simulation

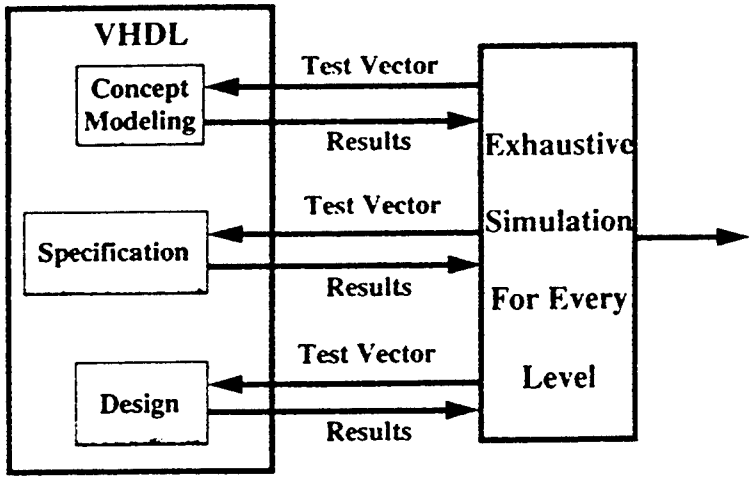


Figure 3: WAVES Test Bench

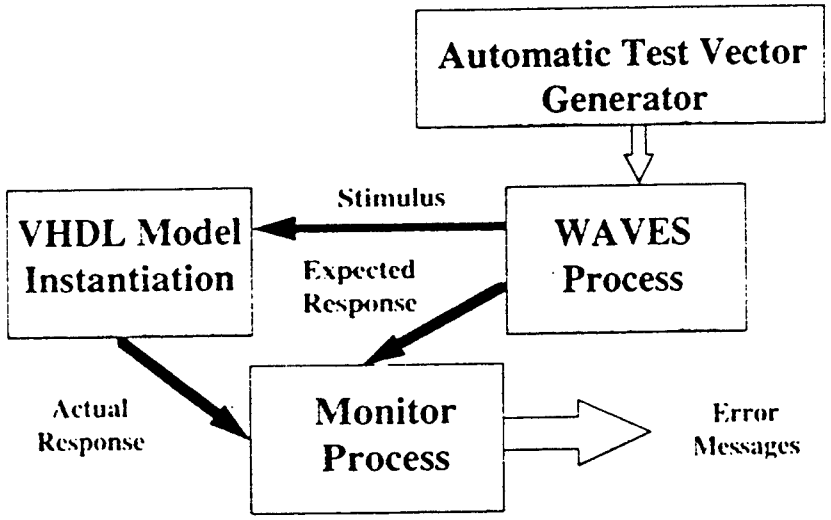


Figure 4: Formal Verification Using Larch/VHDL

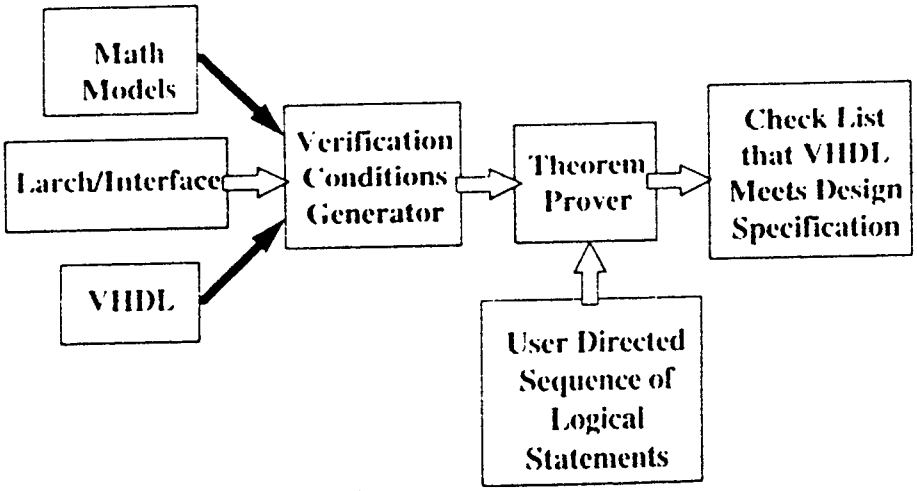


Figure 5: ORA Penelope Theorem Prover

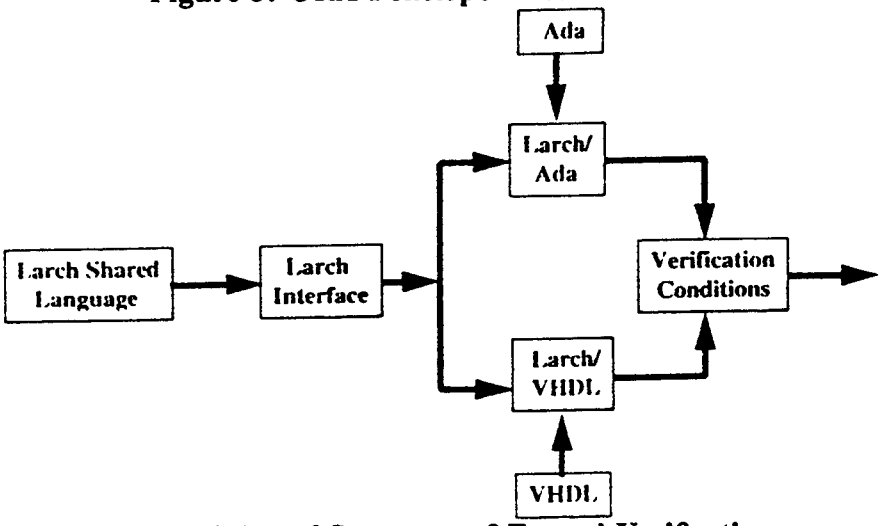


Figure 6: Proof Structure of Formal Verification

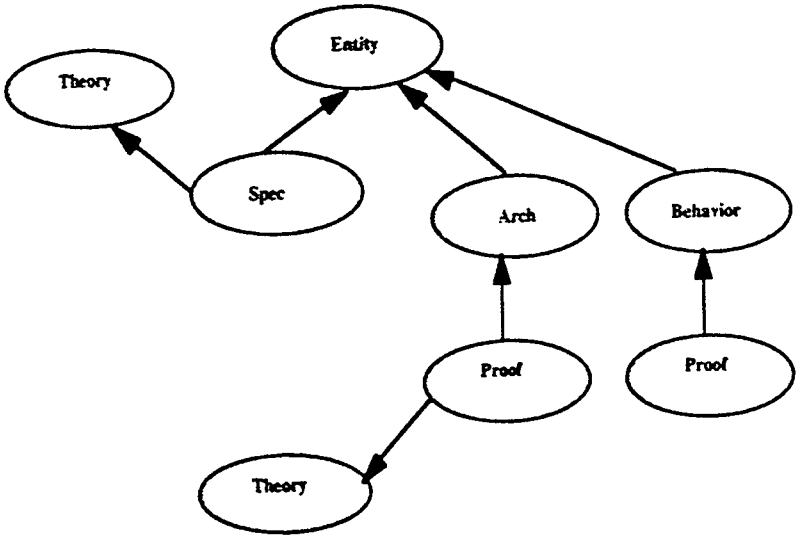


Figure 7: Structure of Four-Bit Ripple Adder with Overflow

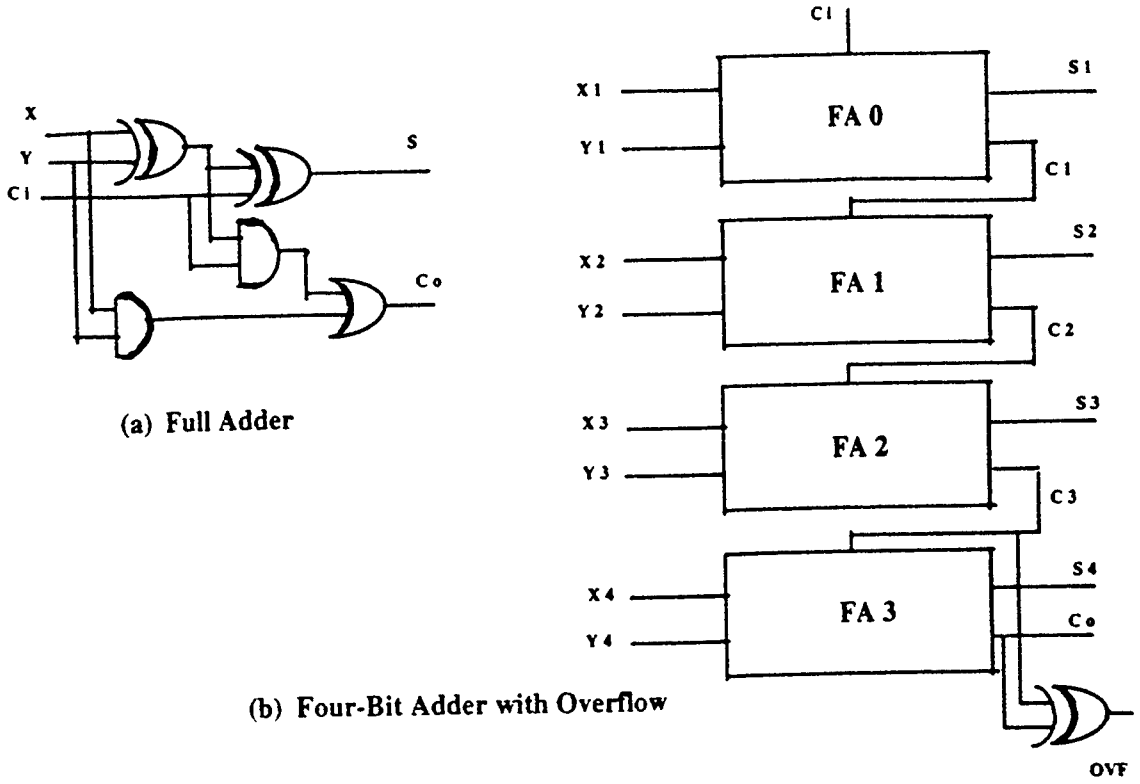
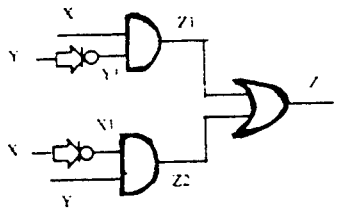
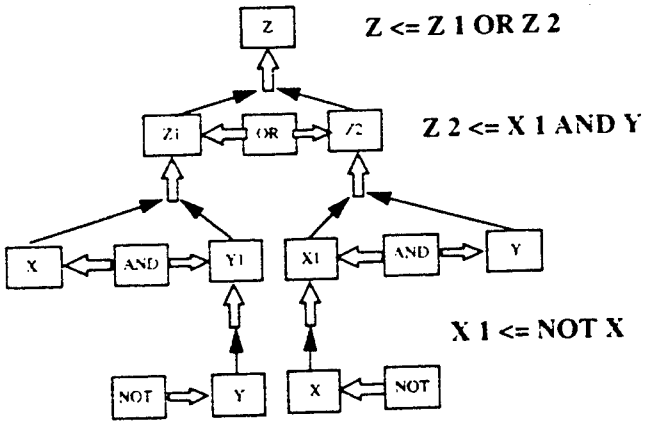


Figure 8: Automated Proof of EX-OR Function



Z <= X XOR Y

Repeated Application of
The Assignment Statement:
Z <= X
And Logical Operators



Formal Verification of EX-OR Function

Z <= (X AND NOT Y) OR (NOT X AND Y)

Figure 9: Testability Conditions for Hazard-Free Circuit

F (X,Y,Z) = SOP (2,3,5,7)

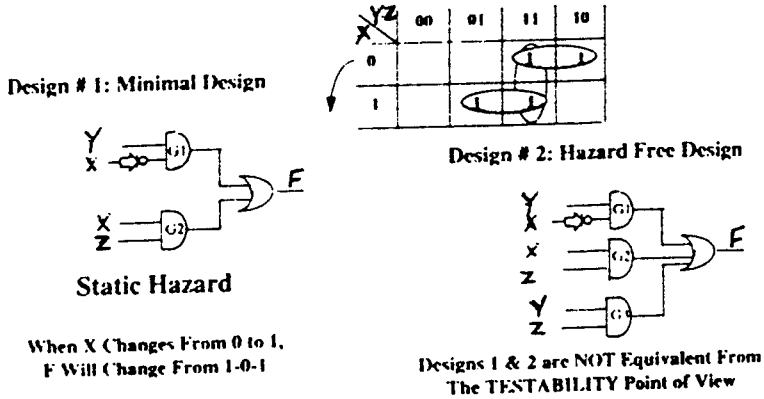
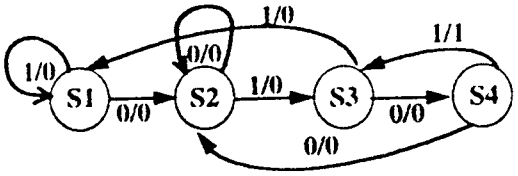


Figure 10: Formal Verification of A Finite State Machine

Next State = F1(Previous State, Input)
Output = F2(Previous State, Input)

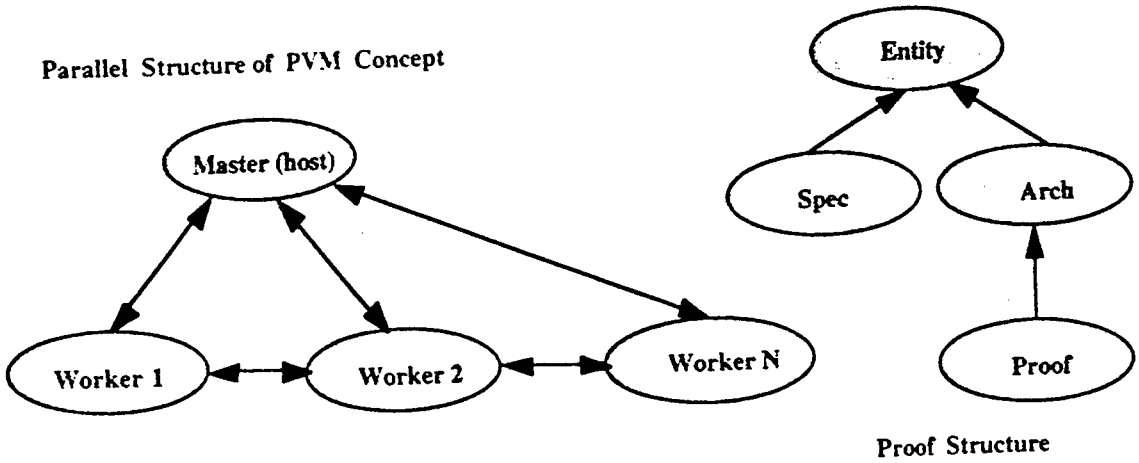


(0101) Sequence Detector
State Diagram

State Table

IN / PS	NS, Z	
	X = 0	X = 1
S1	S2, 0	S1, 0
S2	S2, 0	S3, 0
S3	S4, 0	S1, 0
S4	S2, 0	S3, 1

Figure 11: Structure of Formal Verification vs PVM Parallel Connection



FORMAL SPECIFICATION AND VERIFICATION OF MISSI ARCHITECURE USING SPIN

Milica Barjaktarović
Assistant Professor
Department of Electrical and Computer Engineering

Wilkes University
Stark Learning Center
Wilkes-Barre PA 18766

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force of Scientific Research
Boiling Air Force Base, DC

and

Rome Laboratory

September 1996

FORMAL SPECIFICATION AND VERIFICATION OF MISSI ARCHITECURE USING SPIN

Milica Barjaktarović
Assistant Professor
Department of Electrical and Computer Engineering

Wilkes University
Wilkes-Barre PA 18766

Abstract

In this document we formally specify and verify a part of the Multilevel Information System Security Initiative (MISSI). MISSI is a National Security Agency (NSA) program, designed to send protected messages over unprotected networks such as Internet. MISSI uses several kinds of cryptography for protecting the messages. Cryptography is accomplished using a credit card sized Personal Computer Memory card Interface Association (PCMCIA) card called the FORTEZZA Crypto Card (or "Card", for short).

We constructed a formal specification of sending e-mail using MISSI. We used formal language called Promela, based on Hoare's CSP. We verified the model using automated model checker SPIN, developed by AT&T.

Chapter 1

Introduction

The Multilevel Information System Security Initiative (MISSI) is a National Security Agency (NSA) program, designed to send protected messages over unprotected networks such as Internet [1]. MISSI uses several kinds of cryptography for protecting the messages. Cryptography is accomplished using a credit card sized Personal Computer Memory card Interface Association (PCMCIA) card called the FORTEZZA Crypto Card (or "Card", for short).

MISSI's ancestor is Preliminary Message Security protocol, conceived in October 1991, to process DMS Sensitive But Unclassified (SBU) messages. The name was changed into MOSAIC in 1993. In 1994, the project became MISSI Phase I. Phase I beta version cards were built and distributed to fifty users in April 1992. Currently, MISSI is in Phase II, which includes processing of Sensitive (S) messages. Since FORTEZZA cards can handle only SBU messages, Phase II uses FORTEZZA Plus Cards, which are capable of handling S messages. New release of MISSI will be found in `http://www.armadillo.huntsville.al.us`.

The original system was designed to work with minimum hardware configuration, an Intel 80286 based microprocessor with 512 KB of RAM and operating DOS3.1. MISSI Phase I device drivers are available for DOS, SunOS4.11.3, Solaris 2.3, SCO UNIX, HPUX, and Macintosh platforms as Government Furnished Information (GFI).

Design and testing of a large, complex, multi-million dollar software and hardware system such as MISSI poses many challenges. From the engineering perspective, the two basic questions are:

- how do we build the system?
- how do we know that it works, i.e. is it correct?

In this report, we will discuss the above questions regarding software design, development, and testing. Specifically, we will examine how different components of MISSI interact with each other in order to produce and send an e-mail message.

This work is applicable to any concurrent system design and is based on our work in protocol engineering. The term "protocol engineering" was coined in 1981 in order to label an increasingly important class of problems in the field of computer networks, namely designing communication protocols. Communication protocol is a set of rules used for communication between entities. Protocols are in essence asynchronous software. Protocol design requires unambiguous specification, modularity and step-wise refinement, which may be effectively accomplished by using formal methods for specification, validation and verification.

Protocol designers are especially concerned with testing the consequences of specifications and gaining confidence in their appropriateness (validation), and showing that an implementation satisfies the specification (verification). Verification is one of the key reasons for the interest shown in practical application of formal methods to protocol design, since all different implementations must conform to the specification in order to be mutually compatible. Standards organizations, like the International Standards Organization (ISO), are especially concerned with this issue. ISO has

actually developed and recently standardized two formal methods to be used for specifying their protocol standards.

The basis of our work is application of a mathematically based formal method to specify and verify protocol, in order to gain better understanding of protocol design by formal reasoning and proofs.

The goals of our work are:

- help the designers of complex systems comprehend and design the system in an organized way
- help the communication between teams designing and building various parts of the system
- eventually hand the verified formal specification to the programmers and builders; based on their input, change the specification, and repat this step.

We chose formal language Promela, based on Hoare's CSP, to specify and test sending e-mail using MISSI.

Chapter 2 gives an overview of formal tools, Chapter 3 describes MISSI, Chapter 3.3 presents an example of research done so far and the results, and Chapter 3.4 presents the conclusion and discusses the future research.

Chapter 2

Methodology

In order to understand and formally specify a system, we need to:

- identify components, both software and hardware
- identify the functionality, i.e. behavior of each component and the system (i.e. produce a requirements specification)
- specify the control path (i.e. which steps are to take place)
- specify data path (i.e. what data needs to be stored and exchanged).

Complex and/or concurrent system tend to be too large for analysis by hand. In order to specify, validate and verify a complex system like MISSI, we need to use systematic approaches and automated tools. We have several types of approaches and tools at our disposal:

- rapid prototyping
- simulation
- model checkers
- theorem provers.

Rapid prototyping is a process in which a prototype of the actual system is built. This method can be very useful for designing user interface, especially if it is a graphical user interface (GUI). Users can get acquainted with the system and add/delete functionalities of the system.

Simulation involves building a non-deterministic model of the system and running the model for many test cases. Models can be built in any programming language or in specialized simulation languages like GPSS. Simulation is useful for obtaining performance evaluation, although the results may not be accurate because the users have to estimate all probabilities. In addition, it is impossible to generate all possible test cases, and events with smaller probabilities may never happen.

2.0.1 Model Checkers

We will discuss model checkers used to test process algebra specifications using temporal properties.

Process algebras are model-oriented formal specification languages which specify a system's behavior by constructing a model of the system in terms of mathematical structures such as tuples, functions, sets, and sequences. Other languages that belong to this category include: Parnas' state machines, VDM, Z (used to specify sequential programs and abstract data types), and Petri nets. The most well-known process algebras are Milner's Calculus of Communicating Systems (CCS), and Hoare's Communicating Sequential Processes (CSP), used to specify concurrent programs and distributed systems. Process algebra with no value-passing specifies the control path, and the addition of value-passing provides data path.

Temporal logic is a property-oriented formal language, used to describe properties that a system should have. Therefore, temporal formulae are well-suited for writing the requirements specification of the system, and writing tests to prove that the system satisfies the requirements.

Model checkers are fully automated verification tools which usually work using state exploration. A model checker accepts a process algebra description and tests it using modal or temporal logic or some other mechanism such as assertion statements about data. Therefore, process algebra provides a description of all possible actions taking place in the model. Temporal logic provides a description of properties that event sequences must have. Usually, model checkers have built-in features for testing for absence of deadlock, livelock, and various other properties.

Model checkers are relatively easy to learn and provide relatively quick and useful results. One drawback of these tools is that, since they explore state space, they can run out of memory. Therefore, larger models cannot be modeled in entirety, but we model only selected portions of the system.

Some model checkers work without value-passing, for example Concurrency Workbench (CWB) uses plain CCS. Some model-checkers have value-passing, for example SPIN uses a variant of CSP called Promela. While the addition of value-passing is useful for modeling a system, value-passing adds states and has to be employed selectively for larger models.

Model checkers with value-passing can be thought of "simulators of all possible paths," because the values supplied have to be given finite, specific range.

2.0.2 Theorem Provers

Theorem provers are automated verification tools which prove properties that a system should have. There are first-order and higher-order logic theorem provers, such as Penelope and HOL, respectively. Theorem provers accept logic description of system properties (i.e. axioms), and logic description of desired system properties (i.e. theorems). User has to prove theorems using axioms, manually guiding every step of the proof. Data can have any range. Proofs are constructed by induction for all data values in a given range.

Theorem provers are more involved tools than model checkers and take longer time to learn.

2.1 Tool Selection

Since no single tool can provide a complete view of the system, we ned to combine them for optimal results.

In order to better illustrate the the real-life needs, assume that we are specifying a protocol which uses serial numbers 0-400,000 to mark packets. Model checkers can specify event sequences and test the properties of event sequences, but cannot specify all data to be used. Theorem provers can specify properties valid for all data in a given range but cannot specify event sequences. Therefore, we can combine model checking and theorem proving to provide a more complete specification of the system.

For example, let's assume that we are modeling sending e-mail: issuing an e-meil request, processing text, producing a packet, and sending it to the network. In plain CCS, this specification is:

```
Send_mail = email_req.process_txt.'valid_packet.'send_packet.Send_mail
```

In value-passing CCS, the specification is:

```
Send_mail = email_req.process_txt(txt).
            'send_packet(valid_packet(txt)).Send_mail
```

We can use temporeal logic to verify that every valid e-mail request produces a valid packet, i.e. every time we issue an e-mail request we will eventually have a valid packet:

$Send\_mail \models BOX[email\_req](EVENTUALLY <' valid\_packet > T)$.

In Promela, the above specification can be written as:

```
proctype Send_mail(byte text)
{
  email_req?1;
  process_txt?txt ;
  valid_packet = txt + hdr;
  send_packet!txt ;
}

byte txt = 1;
init() {run Send_mail; email_req!1; process_text!1}
```

Therefore, in process algebra we cannot specify what "valid packet" means, but we can indicate that an event of producing a valid packet must happen. A theorem prover such as HOL can specify all properties of a valid packet, but not at what point those properties hold. Therefore, we can interleave the two specifications and claim that HOL properties of valid packets hold at the instances specified by process algebra specification.

2.1.1 Which Model Checker to Use

We have evaluated several model checkers:

- Concurrency Workbench (CWB)
- Concurrency Factory
- VPAM
- SPIN
- VHDL Penelope

CWB was developed at the University of Edinborough. It accepts CCS specifications, which then can be tested in many ways, including: validation, verification by observation equivalence, testing for deadlock, livelock, safety and liveness, monitoring the behavior of the system by looking at possible state transitions, checking if desired properties hold true, and simulating a desired process step by step. For a detailed description of CWB and specification and verification of a large model, see [2]. Glenn Bruns, currently at AT&T, developed a translator from value-passing CCS into plain CCS. This translator expands value-passing description into non-value passing by translating data into events. Since CWB's capability is sensitive to specification style, i.e. specification requires hiding as soon as possible in order to require less memory space, we feel that the value-passing enhancement needs to address that issue in order to be useful for larger models.

Concurrency Factory was developed by Scot Smolka from SUNY Stony Brook. The Factory addresses the weaknesses of CWB by adding value-passing, more efficient algorithms, and C-like specification style. Eventually, it will be able to produce C code from the specifications. It is still building the user base.

VPAM consists of value-passing CCS embedded in first-order logic theorem prover. Certain CCS laws are provided as "buttons" so that the user does not have to do it manually. We have downloaded VPAM and executed some examples, but we found bugs and lack of documentation. It is highly unfortunate, as the tool is promising.

SPIN was developed by Gerard Holtzman at AT&T. SPIN accepts specification in Promela, which is a variant of CSP. Promela specifications have value-passing and resemble C code. Specifications can be tested for livelock, deadlock, and temporal properties.

Penelope is an automated theorem prover which accepts first-order logic formulae. It has been developed by Odyssey Research Associates (ORA), Ithaca, USA. VHDL is a hardware description language developed by the government, and has been embedded in Penelope. VHDL produces state machine description of a system and is suitable for software applications.

Chapter 3

Overview of MISSI

MISSI provides the following security services:

- data integrity and authentication
- confidentiality
- non-repudiation with proof of origin
- non-repudiation with proof of receipt (optional).

Data integrity means that the data received is the same as the data sent, i.e. it has not been changed in transit. This service is accomplished by hashing and digitally signing messages. Authentication of a message ensures that the signature on the message belongs to an authorized user of MISSI. Users are authorized by certification authorities, which issue certificates ("tickets") to the users.

Confidentiality is accomplished by encrypting data with a secret key. Messages are then digitally signed. MISSI uses a combination of public and secret cryptography. Each message is encrypted using a secret key, and the secret key is e-mailed using public keys.

In order to illustrate differences between secret and public key cryptography, let us assume User A is sending mail to User B. In secret key cryptography, A and B share the same secret key to encrypt and decrypt a message. Therefore, A and B have to somehow exchange the secret key before they can exchange messages. In publc key cryptography, each user has his own private key as well as a public key. A encrypts the message using A's private key x(A) and B's public key y(B). B decrypts the message using A's public key y(A) and B's private key x(B). Therefore, there is no need to exchange any keys, as public keys are publicly available, and each user guards his private key.

Non-repudiation with proof of origin is accomplished by digitally signing outgoing messages.

Non-repudiation with proof of receipt is accomplished by having the receiver provide a signed receipt for incoming messages.

These services depend on the following assumptions:

- both sender and receiver use the same hash and cryptographic algorithms.
- each user has a unique distinguished name
- PAA can be trusted
- user's card stores the proper identification for PAA

3.0.2 How MISSI Works: an Overview

In order to illustrate how MISSI works, we will assume that User A wants to send e-mail to User B.

A will login with his Card, supply his PIN number, write a message, and his Card will encrypt the message using a secret key and sign the message. The secret key is encrypted using public key cryptography and sent in the header of the message. Receiver will extract the secret key using public decryption, and decrypt the message. Receiver will also verify sender's signature, and signatures of authorities that signed sender's signature.

The process of encryption and signing of the Card is described in more detail in figures 3.9, 3.10, and 3.11. Contacting the Card to apply security services to a message involves SDN.701 Message Security protocol (MSP) and is described in more detail in section 3.1. The process of sending out a message to the Internet involves using X.400 or RFC1521 message formatting mechanisms, which will be discussed in section 3.0.3.

Public keys are posted in the MISSI Directory. Each user's private keys and PIN are stored on the user's Card and non-readable by the user. The MISSI Directory is the X.500 Directory equiped with a FORTEZZA card. X.500 Directory is a distributed "white/yellow pages" repository of public keys. We will describe the X.500 Directory in section 3.0.4.

Public keys are stored in **certificates**. A certificate is a data structure which includes each user's name and public keys, signed by the authority that issued the certificate. A certificate consists of: version, serial number, issuer's signature algorithm, issuer's distinguished name validity period, subject's distinguished name, subject's public key information, and issuer's signature. Certificates and keys can expire, get compromised in some way, or be revoked. Authorities keep invalid certificates' serial numbers in Certificate Revocation Lists (CRLs). Authorities also keep Key Revocation Lists (KRLs). X.509 standard specifies the hierarchy of certification authorities and certificate and key management policies. We will discuss X.509 standard in section 3.0.5.

In Figure 3.1 we show the overal MISSI architecure and identify MISSI components. The figure represents an enclave consisting of a Local Area Network (LAN) with MISSI components, attached to the Internet via a Secure Network Server (SNS). We are assuming the enclave is classified as Secret. Sensitive but Unclassified enclaves can have a commercial firefall instead of SNS [3]. Other enclaves are attached to the WAN, but we don't show them to conserve the space.

Workstations are equiped with either FORTEZZA Cards (F) or FORTEZZA Plus cards (FP), or they have no cards. Users are certified by the Certification Authority (CA) operating at the CA workstation (CAW). Audit Manager (AM) is used for auditing the system. Mail List Agent (MLA) is used to forward mail to a list of recipients. Rekey Manager (RKM) is used to rekey the cards.

Directory Server (DS) represents a portion of MISSI Directory. Policy Aproving Authority (PAA) and Policy Creation Authority (PCA) are higher-level certification authorities, operating at PAA workstation (PAAW) and PCA workstation (PCAW).

3.0.3 X.400 Message Handling System

In Figure 3.2 we show the overview of X.400 Message Handling System. X.400 is a suite of standards for message formating and transfer. Each workstation has application software called User Agent (UA), which sends and receives messages. Message Transfer Agents (MTAs) route messages through the Message Transfer System (MTS) (sometimes called Message Handling System (MHS)). Messages can be stored in Message Storage (MS).

3.0.4 X.500 Directory Service

In Figure 3.3 we show the overview of X.500 Directory service. X.500 is a suite of standards for the distributed Directory service. The Directory is a distributed database which contains public information about users, i.e. it relates users' distinguished names with their certificates. The Directory can be used as white or yellow pages. UA contains Directory User Agent (DUA), which contacts the Directory through Directory Service Agent (DSA).

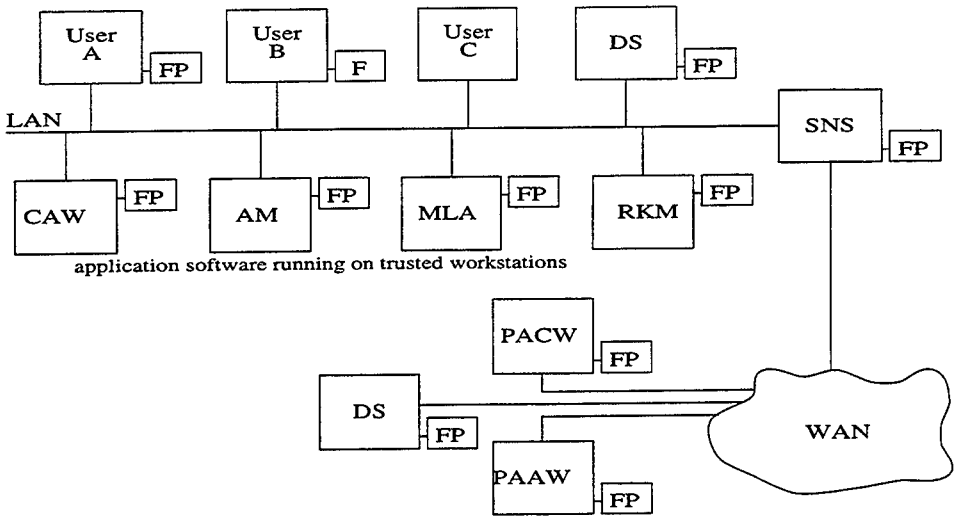


Figure 3.1: MISSI Components

In order to reduce network traffic, users are asked to store verified certificates in a local cashe. If the local cashe does not contain information requested, the local Directory is contacted. There may or may not be a local part Directory on an enclave. Users are permitted to obtain certificates from a floppy [3]. If the local directory cannot find the information requested, it will contact other parts of the DIrectory to obtain the information.

3.0.5 X.509 Certificate and Key Management

X.509 standard specifies the certification hierarchy and certificate and key management.

The abstract view of certification hierarchy is shown in Figure 3.4.

On top of the hierarchy is the Policy Aproving Authority (PAA), which is the trusted authority. (Authentication depends on having at least one trusted source.) PAA's signature is stored on each user's Card, and used for the final authentication. PAA issues certificates for Policy Creation Authorities (PCAs), but cannot revoke any PCA and does not keep Certificate Revocation List (CRL) for PCAs.

Each PCA issues certificates for Certification AUthorities (CAs) in its domain, and maintains a CA CRL. It also maintains a CA KRL, and distributes it to all CAs in its domain. PCA will post signed CRL and KRL to the Directory.

Each CA issues user certificates, and maintains and posts user CRL to the Directory.

Note: the notation changed, from Root Registry to PAA, from Root Authority to PCA, and from Local Authority to CA.

PAA, PCA and CA operate through a CA workstation (CAW) which can be set up to provide PAA, PCA or CA capabilities. CAW has a trusted operating system, and is connected to a FORTEZA device (Figure 3.5. CAW must be able to interface with X.400 and X.500 standards.

3.0.6 MISSI Workstation

In Figure 3.6 we show software layers involved in preparing a MISSI message. These layers are on the application level. First, the message is composed using a message preparation editor. Second, UA prepares the message according to a specified standard. Third, security services are applied to the message using MSP protocol, which contacts the cryptography mechanism located on the

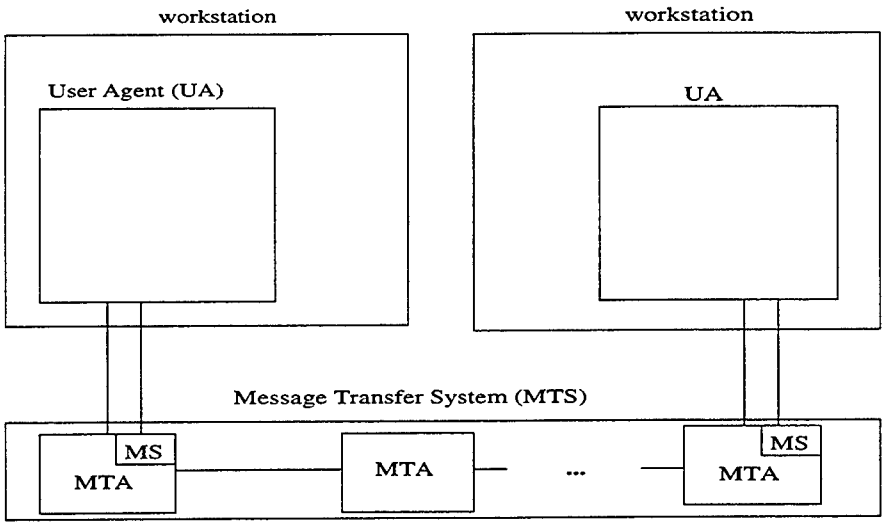


Figure 3.2: X.400 Message Handling System

Card. Finally, the message is sent to the Internet using either X.400 or SMTP/MIME formats. The message is then handed to TCP/IP Internet protocols, which transport the message to the receiver.

In Figure 3.7 we show the abstract view of the interface between a MISSI workstation and FORTEZZA card. UA application process contacts the hardware on the Card through MSP application software. MSP application software represents a high-level interface to the Card and consists of a library of eight `msp_` functions. These functions call on the library of fifty one Crypto Interface (CI) `CI_` functions, which interface with the card device driver and eventually with the hardware. The Card's hardware includes a CAPSTONE chip which performs cryptographic functions, and volatile and non-volatile memory.

In Figure 3.8 we show the software components within a MISSI workstation and its connection to the FORTEZZA hardware.

3.1 SDN.701 MSP

MSP interface with the card includes the library of the following functions:

msp_login	manages FORTEZZA card login
msp_clear	loggs the user off the card
msp_submit	builds signed and/or encrypted MSP messages
msp_status	pre-processes an MSP message
msp_deliver	decrypts and validates MSP messages
msp_check_sig	verifies signature of block of data
msp_val_receipt	validates a signed receipt
msp_mla_proc	provides security services to MLA.

Use of `msp_login` function is illustrated in Fig. 3.12. Use of `msp_check_sig` and `msp_submit` functions is illustrated in Fig. 3.13 and Fig. 3.14.

The certificates are verified following the procedures outlined in SDN.701. In order to validate a certificate, UA uses bottom-up approach: it obtains all certificates and CRLs up the certification chain (e.g. user, CA, PCA, PAA) until the root of the chain (PAA) is reached.

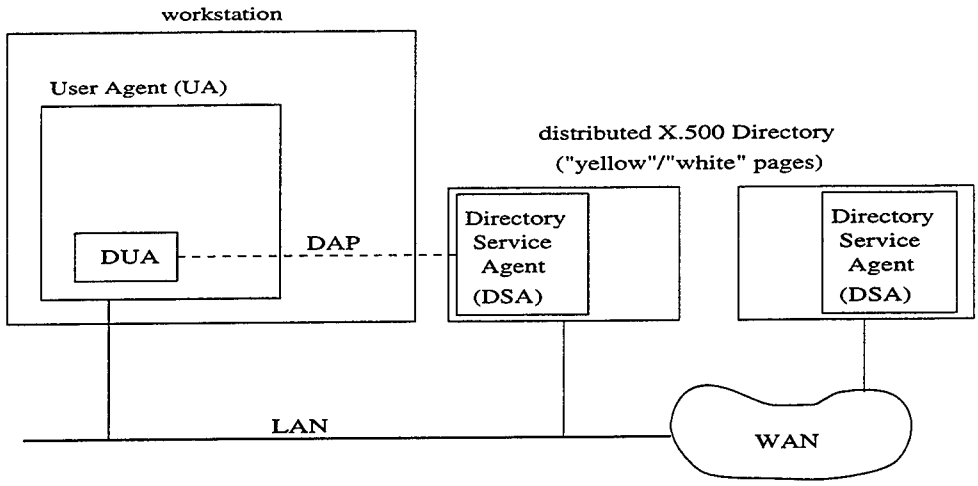


Figure 3.3: X.500 Directory Service

Once PAA's certificate is obtained, certificates are verified top to bottom, starting with PCA's certificate. Broadly speaking, each certificate is verified according to the following procedure:

- validate issuer's signature on certificate
- ensure certificate does not appear on revocation list
- validate certificate privilege fields
- validate that certificate subject is subordinate to issuer

The above procedure is repeated for all certificates down the authentication chain. The FORTEZZA card is used to do the signature checks.

PAA's signature is always checked using PAA's certificate stored on the Card.

3.2 Secure Network Server (SNS)

Enclave users send certificate requests to the local part of the Directory (i.e. the local DSA), and they obtain certificates from the local DSA. X.500 states that, if the local directory cannot find information requested, it queries other DSAs on the network to find the information. However, the current version of SNS will not allow X.500 inquires to leave or enter enclave. (This feature may be changed in the future.) The enclave is configured to rely on internal directory information. The DSA within the enclave is not be configured to "know" about external DSAs. Users requesting information that does not exist in the internal DSA's database will receive an "information does not exist" response. (Even if the local DSA knew about the external DSAs, the connection between them would not be allowed through the SNS. It is assumed that some external to internal replication mechanism is provided for the local DSA.)

Therefore, if the local DSA does not contain the certificate being asked for, the user cannot send any message, since SNS will not let the user send a request to the WAN.

SNS will verify certification path in each MISSI message that contains a certificate, either for messages from enclave or to the enclave. The SNS will have the ability to query external DASs for certificate information needed to verify the certificate path. X.500 traffic is forbidden only from network to network across the SNS, not from within the SNS out to a network.

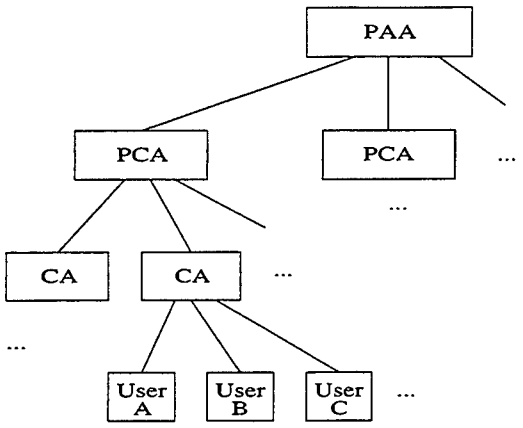


Figure 3.4: X.509 Certification Hierarchy: Abstract View

If SNS receives a message containing a certificate, it verifies the certificate, going all the way up to the "trusted" (i.e. PAA) certificate. The SNS FORTEZZA card contains the complete certification path for the SNS certificate.

3.3 SPIN Model of MISSI Architecture

Our model describes steps needed to send mail using MISSI. We focus on Fig. 3.7. Our model represents Fig. 3.8, with capabilities represented in Fig. 3.12 and Fig. 3.13. Certificate verification is outlined in section 3.1.

The assumptions are:

- users are registered
- cards are initialized
- users are ready to send or receive email.

In the future models, user registration and card initialization process can be added. We will also add receiving capabilities, SNS, and WAN.

The tool we chose for this project is SPIN.

We verified that:

- `msp_` functions are never given PAA's certificate, but always use PAA's certificate from the Card.
- `msp_login` is always called before `msp_submit`.

3.4 Conclusion

Areas of additional research include:

– key generation

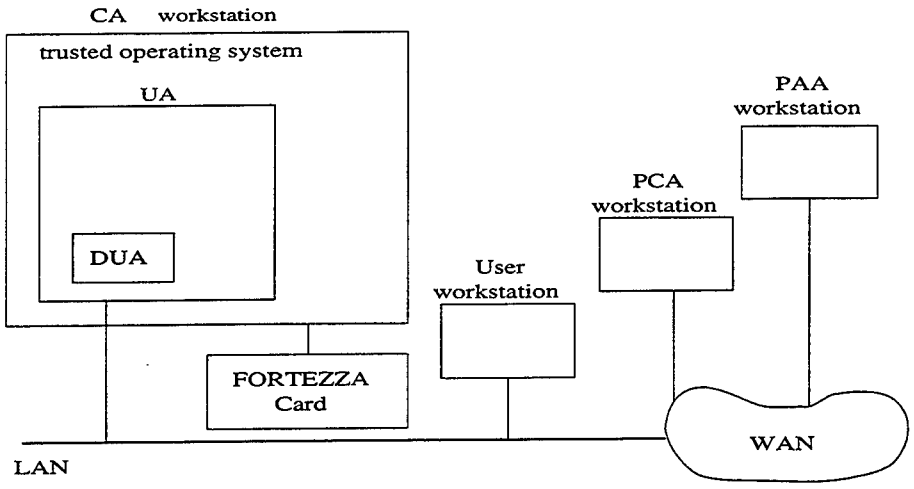


Figure 3.5: X.509 Certificate and Key Management

  - key distribution
  - key replacement (e.g. archiving, revoking keys)

- certificate hierarchy of PAA, PCA, CAW
- adding a new PAC or CAW or user
- testing scenarios

Recommendations for improving model-checkers:

- implement more efficient state-exploration algorithms
- automatically produce $C/C++$ code from the specification
- do not abandon new tools shortly after they are developed; we need robust, mature, well-used tools

Finally, we conclude that:

- Formal specifications lead to a clear description of the system, and aid in understanding of the system and expose weak areas of system design.

  For example, we found that certificate hierarchy is not well designed in MISSI.

- Formal verification assures of desired system properties.

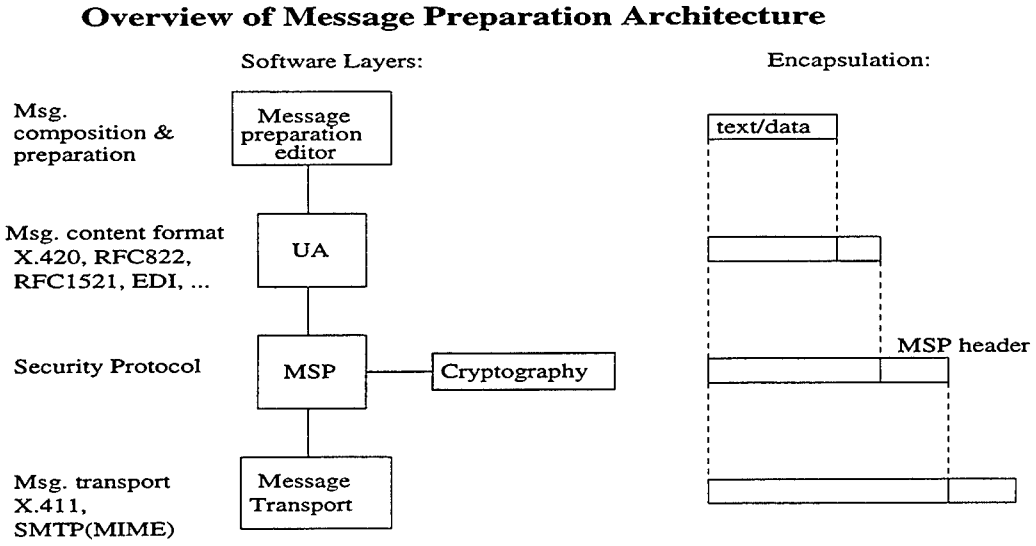


Figure 3.6: Overview of Message Preparation Architecture

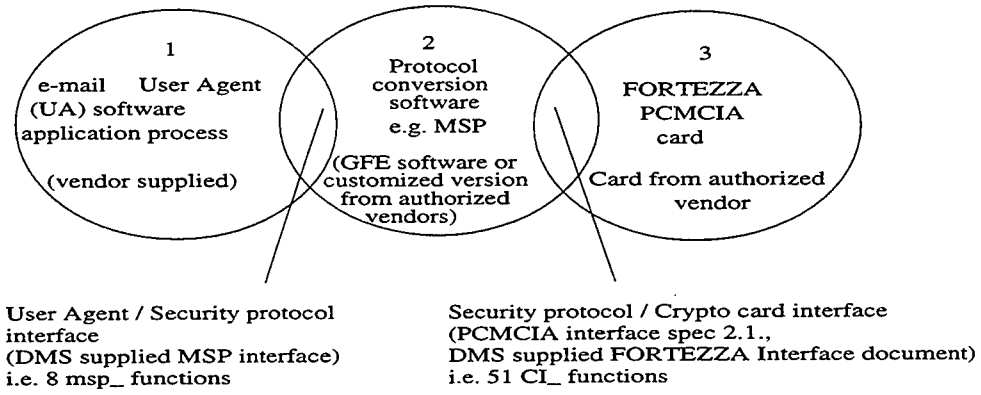


Figure 3.7: Abstract View of MISSI Workstation /FORTEZZA Card Interface

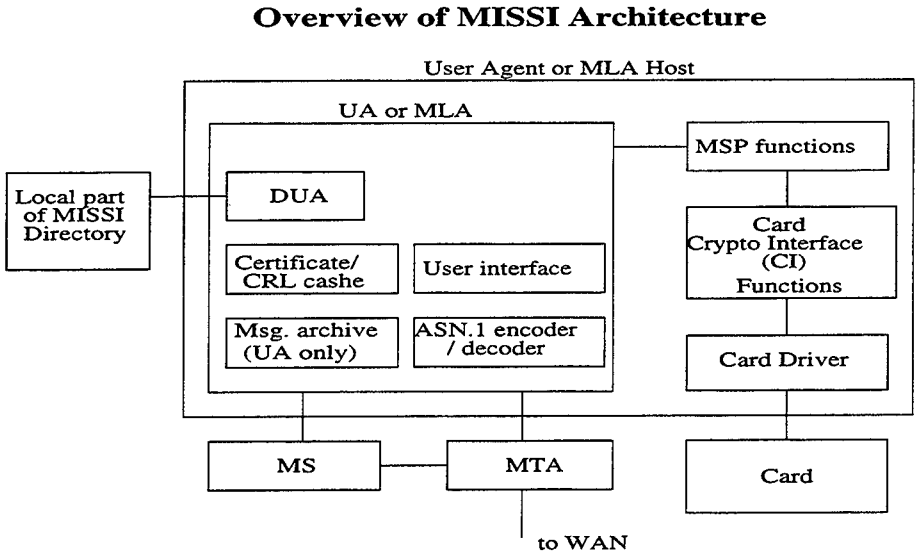


Figure 3.8: Local Workstation Components

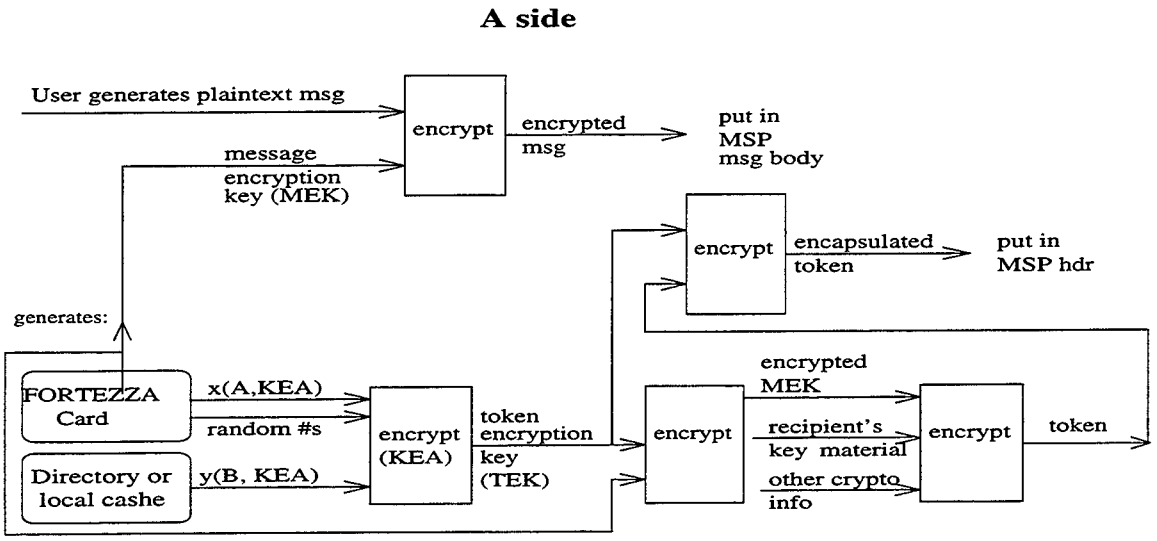


Figure 3.9: Encryption on the Sender's Side

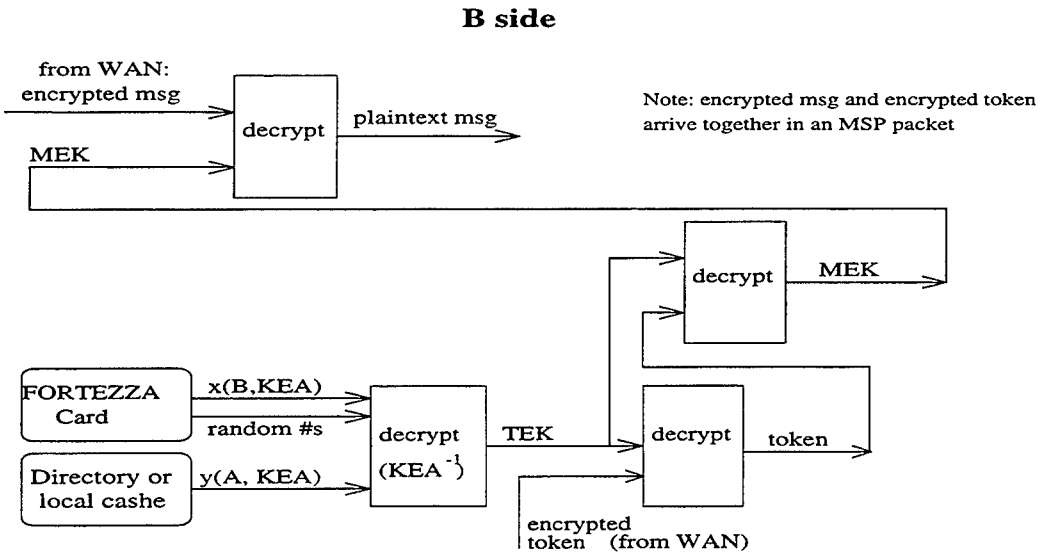


Figure 3.10: Decryption on the Receiver's Side

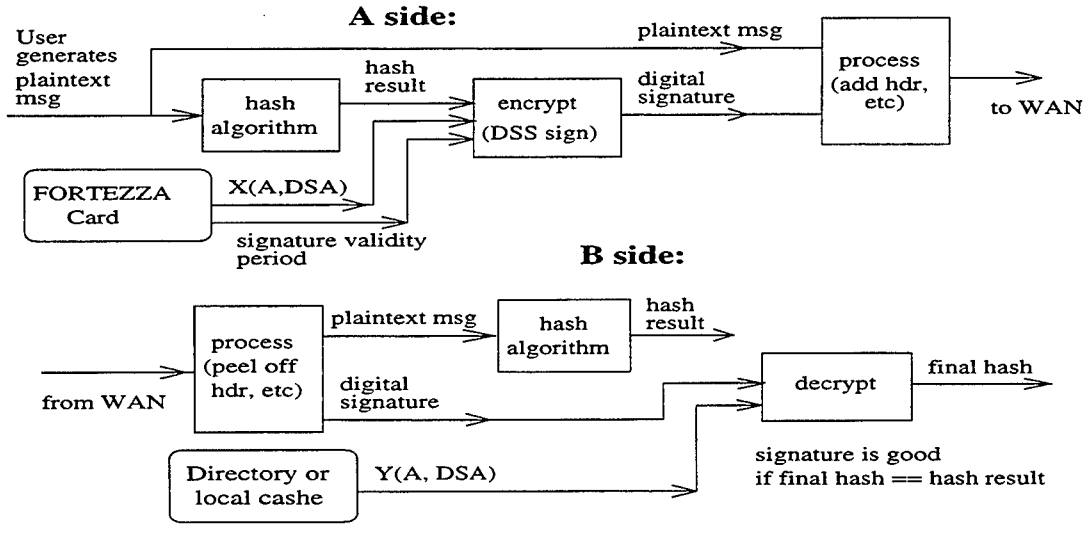


Figure 3.11: Signature Generation and Verification

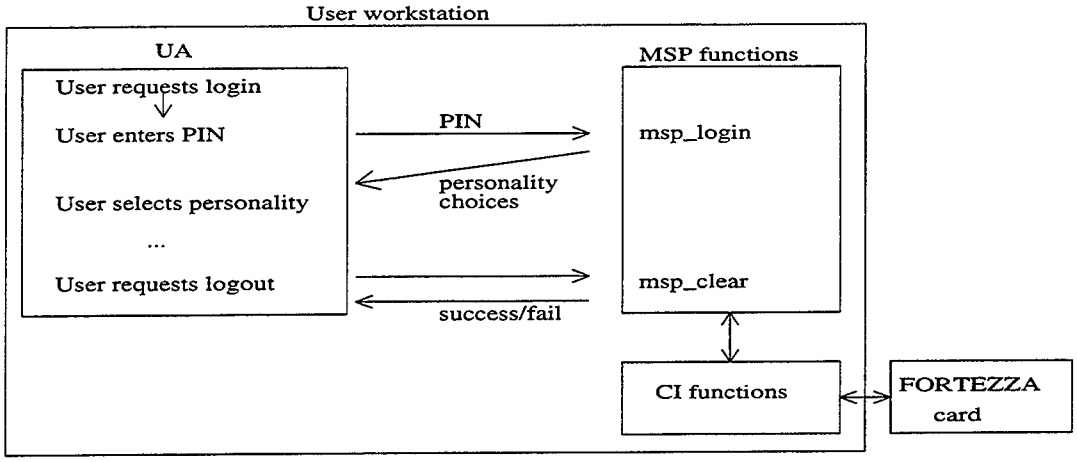


Figure 3.12: MISSI Login

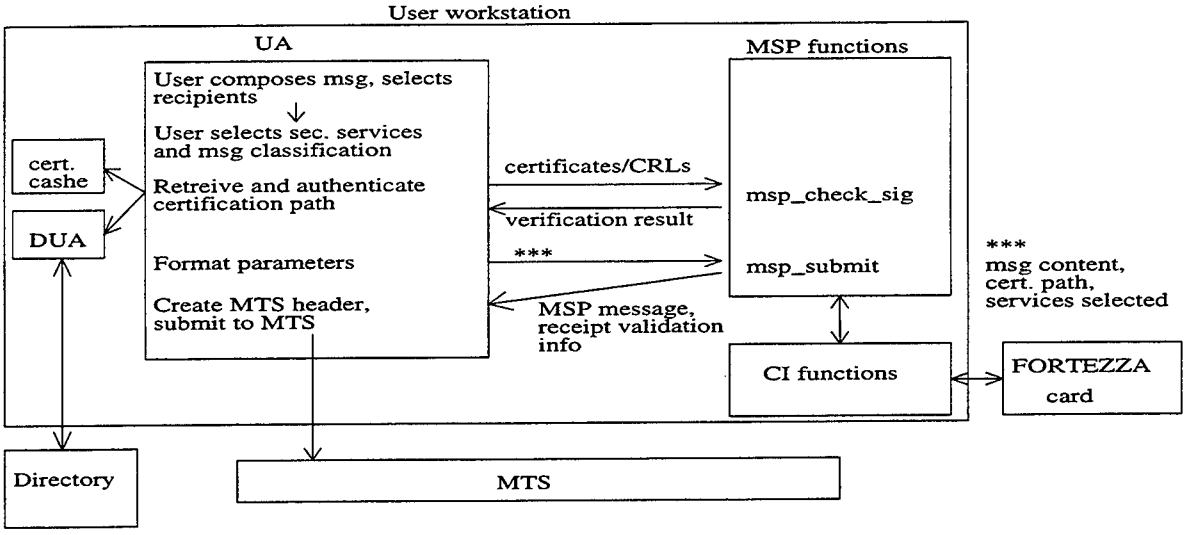


Figure 3.13: MISSI Send Mail

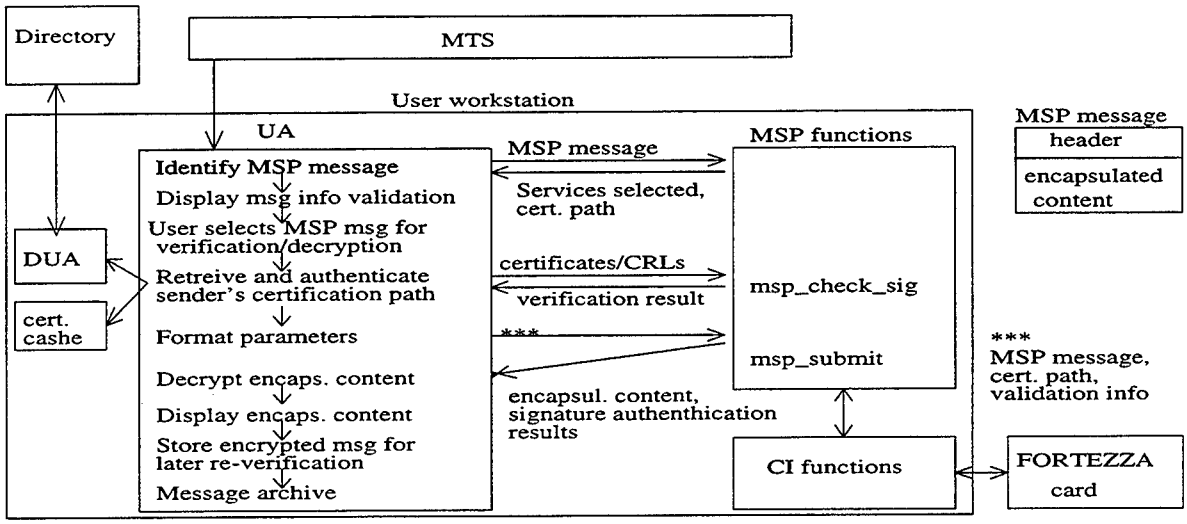


Figure 3.14: MISSI Receive Mail

Bibliography

[1] *FORTEZZA Application Developers Guide.* http://www.armadillo.huntsville.al.us, obsolete, 14 July 1995.

[2] Milica Barjaktarović. *Formal Specification and Verification of the OSI Session Layer Using CCS.* PhD thesis, Syracuse University, http://wilkes1.wilkes.edu/ barjakta, July 1995.

[3] National Security Agency, http://www.armadillo.huntsville.al.us. *Network Security Managers (NSM) Functional Requirements Specification and Concept of Operations (CONOP)*, June 1996.

PERFORMANCE ANALYSIS AND SIMULATION RESULTS OF DELAY AND MULTIPLY RECEIVERS PROCESSING A SPREAD SPECTRUM MODULATED FLIP WAVE-SIGNAL GENERATED FROM HIGH BANDWIDTH EFFICIENCY PULSES

Dr. Daniel C. Bukofzer, Professor and Chairman
Department of Electrical and Computer Engineering
California State University
Fresno California 93740-0094

Final Report for:

Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC
and
Rome Laboratory

September 1996

PERFORMANCE ANALYSIS AND SIMULATION RESULTS OF DELAY AND MULTIPLY RECEIVERS PROCESSING A SPREAD SPECTRUM MODULATED FLIP-WAVE SIGNAL GENERATED FROM HIGH BANDWIDTH EFFICIENCY PULSES

Dr. Daniel C. Bukofzer, Professor and Chairman
Department of Electrical and Computer Engineering
California State University
Fresno California 93740-0094

Abstract

This report presents performance results of delay and (complex) multiply (D & M) receivers processing a specific spread spectrum modulated digital communication signal. The signal structure involves various types of high bandwidth efficiency (HBE) pulses that affect the spectrum of the output generated by the D & M receivers. Although mathematical models and analytical methods have been used to describe and partly set up the problem solution, the performance results of interest are obtained strictly via computer methods utilizing block oriented simulation software. More specifically, the signal of interest, having a unique structure, was mathematically described and generated by the simulation software. This so-called flip-wave signal is generated by a random data signal whose spectrum is spread using pseudo-noise (binary) codes. Much of this effort involved a determination of the detectability of the spectrally spread signal by D & M receivers as a function of different waveshapes and modulation of the amplitude of the waveforms associated with the pseudo-noise codes. The different waveshapes considered included rectangular, raised cosine, and sinc function pulses. Amplitude modulation involving these waveforms was of discrete type that could involve four, six, or up to eight levels. The results demonstrated in general (but not in every specific case), that amplitude modulation and sinc function shaping tended to result in a signal that was more difficult to detect by a D & M receiver. Under such conditions, the signal constructed appeared as nearly featureless. Furthermore, since two receivers were considered, namely the baseband and the carrier D & M device, it was found in essentially every case, that the latter was not as effective as the former, in identifying the presence of the signal under study given the spectral strength of its output at critical frequencies. Not all possible cases and operational scenarios could be considered in this effort. Therefore, in a certain sense, the work completed as part of this project, can be viewed as the starting point for more sophisticated performance evaluation efforts in which a signal detection algorithm is considered as operating jointly with the D & M receiver and practical effects, such as noise, co-channel interference and the effect of multipath propagation are included in studies and system simulations.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

This work was supported by the U.S. Air Force Rome Laboratory, located in Rome, New York, under the auspices of the Air Force Office of Scientific Research, Bolling AFB.

At Rome Laboratory, the author gratefully acknowledges the input provided and time spent during the many technical discussions by Mr. John Patti. His support and assistance proved invaluable insofar as progress on this project is concerned. Additionally, Mr. William Cook is to be thanked for assigning this interesting and challenging project and

providing the support necessary for its completion. Mr. Richard Smith provided much needed assistance with the use of the workstation on which the simulations were carried out. His help is very much appreciated. Many thanks to Mr. Steve Yax for providing the administrative leadership, support, and flexibility that were needed to complete this project successfully and in a timely manner.

A sincere thank you to Mr. Scott Licoscos, Ms. Johnetta Thompson, and Ms. Rebecca Kelly of RDL Inc. who assisted with the administrative requirements of this contract.

The author recognizes and thanks the commanders of the U.S. Air Force Office of Scientific Research as well as the U.S. Air Force Materiel Command and Rome Laboratory for supporting the Summer Faculty Research Program which allows university faculty members the opportunity to become involved in projects of importance to the U.S. Air Force. This program provides a mechanism for faculty members to remain technically active and contribute toward solving significant problems of interest to the U.S. Air Force.

INTRODUCTION

The development of direct sequence spread spectrum (DSSS) modulation techniques for the purpose of addressing the unique needs of military communications had its beginnings in the early 1950's, [1]. The principles of DSSS modulation (and demodulation) are thoroughly explained in many textbooks (see [2, 3] for example), and the technology has matured to the point that commercial applications are to be found, for example, in cellular communication systems, position location devices, and multi-user, multi-channel data communication networks, [4]. Due to the inherent low probability of intercept (LPI) feature of DSSS modulated signals, a great deal of effort has been expended in the development of systems and techniques that are able to detect and extract features of these LPI signals. Traditional methods have focused on radiometers for detection of DSSS modulated signals [5], however in the recent past, as a result of the work reported on in [6], delay and multiply (D & M) receivers have been proposed and studied as alternatives to carrying out the signal detection and feature extraction operations, [7, 8]. In the last few years, much attention has been paid to the cyclostationary nature of all modulated signals which more specifically has been focused on developing cyclic feature extractors that have the potential of enhancing the ability to detect, direction find (DF), and identify parameters of DSSS modulated signals, [9, 10]. As a result of this, an almost parallel effort has been ongoing, in which developing more complex waveform designs that make the DSSS modulated signal more difficult to intercept and identify its features, has been the main focus, [11, 12, 13]. The development of the flip-wave signal (FWS) generated from high bandwidth efficiency (HBE) pulses represents yet another effort in this direction that has the potential of producing a near featureless signal that is not easily detected by a D & M receiver, yet exhibits good performance in terms of bit error rate (BER) by a cooperating receiver, [14, 15].

Section 1: The Flip-Wave Signal

Let Z(t) be a random data signal mathematically described by

$$Z(t) = \sum_{k=-\infty}^{\infty} A_k p(t - kT_b) \qquad -\infty < t < \infty \tag{1.1}$$

where the sequence $\{A_k\}$ of ± 1 valued random variables (r.v.'s) represents the data and p(t) is a deterministic pulse. It is assumed that

$$E\{A_k\} = 0 \quad \forall k \qquad \text{(i.e., r.v.'s are zero mean)} \tag{1.2}$$

$$E\{A_k A_n\} = \begin{cases} 1 & k = n \\ 0 & k \neq n \end{cases} \quad \text{(i.e., r.v.'s are independent and identically distributed)} \tag{1.3}$$

and

$$p(t) = \begin{cases} 1 & 0 < t < T_b \\ 0 & \text{otherwise} \end{cases} \tag{1.4}$$

The well-known quadrature phase shift keyed (QPSK) modulation signal can be generated by defining two related baseband signals, that is

$$Z_I(t) = \sum_{k=-\infty}^{\infty} A_{2k} p'(t - 2kT_b) = \sum_{k=-\infty}^{\infty} A_{2k} p\left(\frac{t}{2} - kT_b\right) \tag{1.5}$$

$$Z_Q(t) = \sum_{k=-\infty}^{\infty} A_{2k+1} p'(t - (2k+1)T_b) = \sum_{k=-\infty}^{\infty} A_{2k} p\left(\frac{t}{2} - (k+\frac{1}{2})T_b\right) \tag{1.6}$$

where

$$p'(t) \equiv p(t/2) \tag{1.7}$$

A so-called offset QPSK (OQPSK) modulated signal results when amplitude modulation of sine and cosine carriers is imposed. That is,

$$\begin{aligned} S_{OQPSK}(t) &= A[Z_I(t)\cos 2\pi f_c t - Z_Q(t)\sin 2\pi f_c t] \\ &= A\cos[2\pi f_c t + \phi(t)] \end{aligned} \tag{1.8}$$

where A is a deterministic amplitude, f_c is the carrier frequency and T_b^{-1} is the transmission bit rate, or equivalently $T_b^{-1}/2$ is the symbol rate, (2 bits per symbol). Observe that,

$$\phi(t) = \tan^{-1}\frac{Z_Q(t)}{Z_I(t)} \tag{1.9}$$

so that over any time interval $kT_b < t < (k+1)T_b$, $k = 0, \pm 1, \pm 2, \ldots$, the phase modulation $\phi(t)$ can take on values $\pm\pi/4$, $\pm 3\pi/4$, as a result of the fact that $Z_I(t)$ and $Z_Q(t)$ are bi-polar waveforms. (In the case where these are unipolar waveforms, the values that $\phi(t)$ could take on would be $0, \pm\pi/2, \pi$.)

A related but different waveform that also involves a sinusoidal carrier with one of four possible phase values can be generated in such a way that when $A_k = 1$, the carrier phase in the interval $(k+1)T_b < t < (k+2)T_b$ is that of the carrier in the interval $kT_b < t < (k+1)T_b$ incremented by $\pi/2$ radians. Conversely, when $A_k = -1$, the carrier phase in the interval $(k+1)T_b < t < (k+2)T_b$ is that of the carrier in the interval $kT_b < t < (k+1)T_b$ incremented by $-\pi/2$ radians. This can be translated into a

mathematical description as follows. Define two new sequences of r.v.'s, $\{X_k\}$ and $\{Y_k\}$ which are related to $\{A_k\}$ by the recursions

$$X_{k+1} = -A_k Y_k \qquad Y_{k+1} = A_k X_k \qquad k = 0, 1, 2, \ldots \tag{1.10}$$

Observe that an initial starting time t = 0 is implicit where it is assumed that

$$X_0 = 1 \qquad Y_0 = 0 \tag{1.11}$$

Define now X(t) and Y(t), where

$$X(t) = \sum_{k=0}^{\infty} X_k p(t - kT_b) \qquad Y(t) = \sum_{k=0}^{\infty} Y_k p(t - kT_b) \tag{1.12}$$

Observe that the r.v. A_k affects the waveforms X(t) and Y(t) in the time interval $(k+1)T_b < t < (k+2)T_b$. Moreover, these waveforms take on values 0, ± 1 in such a way that in any time interval $kT_b < t < (k+1)T_b$ if X(t) equals ± 1, then Y(t) must be zero valued, and vice versa. Forming now

$$S_{FPPSK}(t) = A[X(t)\cos 2\pi f_c t + Y(t)\sin 2\pi f_c t] \tag{1.13}$$

a four phase (differential) binary phase shift keyed (BPSK) modulation signal has been produced, hence the subscript in Eq. 1.13. Now, if the pulse duration in X(t) and Y(t) is extended over $2T_b$ sec. long intervals, this produces

$$X'(t) = \sum_{k=0}^{\infty} X_{2k} p'(t - 2kT_b) \qquad Y'(t) = \sum_{k=0}^{\infty} Y_{2k+1} p'(t - (2k+1)T_b) \tag{1.14}$$

These two signals are the basis of the flip-wave signal which in complex baseband representation, takes on the form

$$S'_{FW}(t) = X'(t) + jY'(t) \tag{1.15}$$

The notion of a flip-wave appears to be the result of imposing the data recursion described in Eq. 1.10 in such a way that the transmitted RF signal (without spread spectrum modulation) and described by

$$S_{FW}(t) = \mathrm{Re}\{AS'_{FW}(t)e^{j2\pi f_c t}\} \tag{1.16}$$

introduces differential phase modulation in an OQPSK format where at any one signaling interval, a phase change of ± π/2 rad. must take place (i.e., phase flipping).

The introduction of DSSS modulation can be accomplished in the usual manner by generating spreading codes $c_1(t)$ and $c_2(t)$ using linear feedback shift register (LFSR) systems, where

$$c_1(t) = \sum_{m=-\infty}^{\infty} \alpha_m h(t - mT) \qquad c_2(t) = \sum_{\ell=-\infty}^{\infty} \beta_\ell h(t - \ell T) \tag{1.17}$$

and $\{\alpha_m\}$ as well as $\{\beta_\ell\}$ are $\pm$ 1 valued M-sequences, T^{-1} is the chip rate of the spreading codes, and h(t) is a yet to be specified HBE pulse that is not necessarily restricted to the time interval $0 < t < T$. In its simplest form, h(t) is a rectangular pulse of the form described by Eq. 1.4 (where T_b is replaced with T).

The complex baseband flip-wave spread signal thus generated is described by

$$S'_{FWS}(t) = X'(t)c_1(t) + jY'(t)c_2(t - T/2) \tag{1.18}$$

and the transmitted RF signal becomes

$$S_{FWS}(t) = \mathrm{Re}\{AS'_{FWS}(t)e^{j2\pi f_c t}\} \tag{1.19}$$

The focus of this report centers on the baseband processing of the spread flip wave signal. Hence of major concern is the signal described by Eq. 1.18 where attention should be paid to the fact that the second spreading code $c_2(t)$ is displaced in time with respect to the first spreading code $c_1(t)$ by T/2. Thus, from Eqs. 1.14 and 1.17,

$$S'_{FWS}(t) = \sum_{k=0}^{\infty} X_{2k} p'(t - 2kT_b) \sum_{m=-\infty}^{\infty} \alpha_m h(t - mT) + j\sum_{k=0}^{\infty} Y_{2k+1} p'(t - (2k+1)T_b) \sum_{\ell=0}^{\infty} \beta_\ell h(t - (2\ell+1)T/2) \tag{1.20}$$

Analysis of this signal, in spite of its relative simple structure, is difficult to carry out due to the random nature of the data sequence that gives rise to the sequences $\{X_k\}$ and $\{Y_k\}$. These sequences bear no statistical relationship to the pseudorandom sequences $\{\alpha_m\}$ and $\{\beta_\ell\}$, therefore a simplifying assumption can be made provided that $T << T_b$. That is, if it can be assumed that many chips are contained in a bit interval, then $S'_{FWS}(t)$ can be approximated by

$$S'_{FWS}(t) = X\sum_{m=-\infty}^{\infty} \alpha_m h(t - mT) + jY\sum_{\ell=-\infty}^{\infty} \beta_\ell h(t - (2\ell+1)T/2) \tag{1.21}$$

where X and Y are treated as constants that can take on values $\pm$ 1. It is this signal form that is further analyzed in terms of its being processed by a Delay and (Complex) Multiply receiver.

Section 2: The Delay and (Complex) Multiply Receiver

Two different Delay and (Complex) Multiply (D & M) receiver structures are recognized insofar as this report is concerned, namely the carrier D & M and the baseband D & M processors. If (for example) the input signal to these processors is the signal of Eq. 1.21, then the output signals are given by

$$V'_{carr}(t) = S'_{FWS}(t)S'_{FWS}(t-d) = Re\{V'_{carr}(t,d)\} + jIm\{V'_{carr}(t,d)\} \tag{2.1}$$

and

$$V'_{bsb}(t) = S'_{FWS}(t)S'^{*}_{FWS}(t-d) = Re\{V'_{bsb}(t,d)\} + jIm\{V'_{bsb}(t,d)\} \tag{2.2}$$

for the carrier and baseband D & M receivers, respectively. In these equations, the parameter d represents the receiver delay and the superscript * implies a complex conjugation operation. From Eqs. 1.21, 2.1, and 2.2 obtain

$$Re\{V'_{carr}(t,d)\} = \tag{2.3}$$
$$X^2 \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\alpha_m\alpha_n h(t-mT_c)h(t-d-nT_c) - Y^2 \sum_{\ell=-\infty}^{\infty}\sum_{q=-\infty}^{\infty}\beta_\ell\beta_q h(t-(2\ell+1)T_c/2)h(t-d-(2q+1)T_c/2))$$

$$Im\{V'_{carr}(t,d)\} = \tag{2.4}$$
$$XY[\sum_{\ell=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\beta_\ell\alpha_n h(t-(2\ell+1)T_c/2)h(t-d-nT_c) + \sum_{m=-\infty}^{\infty}\sum_{q=-\infty}^{\infty}\alpha_m\beta_q h(t-mT_c)h(t-d-(2q+1)T_c/2))]$$

and

$$Re\{V'_{bsb}(t,d)\} = \tag{2.5}$$
$$X^2 \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\alpha_m\alpha_n h(t-mT_c)h(t-d-nT_c) + Y^2 \sum_{\ell=-\infty}^{\infty}\sum_{q=-\infty}^{\infty}\beta_\ell\beta_q h(t-(2\ell+1)T_c/2)h(t-d-(2q+1)T_c/2))$$

$$Im\{V'_{bsb}(t,d)\} = \tag{2.6}$$
$$-XY[\sum_{\ell=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\beta_\ell\alpha_n h(t-(2\ell+1)T_c/2)h(t-d-nT_c) + \sum_{m=-\infty}^{\infty}\sum_{q=-\infty}^{\infty}\alpha_m\beta_q h(t-mT_c)h(t-d-(2q+1)T_c/2))]$$

In these equations, the real terms are of particular interest since

$$Re\{\bullet\} = \sum_{m=-\infty}^{\infty}\alpha_m^2 h(t-mT_c)h(t-d-mT_c) + \sum_{\substack{m=-\infty \\ m\neq n}}^{\infty}\sum_{n=-\infty}^{\infty}\alpha_m\alpha_n h(t-mT_c)h(t-d-nT_c) \pm$$
$$\sum_{\ell=-\infty}^{\infty}\beta_\ell^2 h(t-(2\ell+1)T_c/2)h(t-d-(2\ell+1)T_c/2) + \sum_{\substack{\ell=-\infty \\ \ell\neq q}}^{\infty}\sum_{q=-\infty}^{\infty}\beta_\ell\beta_q h(t-(2\ell+1)T_c/2)h(t-d-(2q+1)T_c/2)) \tag{2.7}$$

where the positive sign in Eq. 2.7 corresponds to the baseband signal given by Eq. 2.5 and the negative sign corresponds to the carrier signal given by Eq. 2.3. Observe that a periodic component is always present in Eq. 2.7 that gives rise to a spectral line in the frequency domain representation of this signal whose strength is dependent on the statistics of the sequences $\{\alpha_m\}$and $\{\beta_\ell\}$, the HBE pulse shapes h(t), and the receiver delay setting d. If the two sequences are made up of ± 1 (equally likely) valued components, the periodic components in Eq. 2.7 (first and third terms) are deterministic so that a statistical description is not necessary. These periodic components are very similar in form since they are both functions of M-sequences and the same HBE pulse h(t). Therefore further analysis is carried out only on the first term of Eq. 2.7. First, define

$$\Psi(t) = \sum_{m=-\infty}^{\infty} \alpha_m^2 h(t - mT_c) h(t - d - mT_c) \tag{2.8}$$

and evaluate next the statistical expectation of this periodic process, namely

$$E\{\Psi(t)\} = \sum_{m=-\infty}^{\infty} E\{\alpha_m^2\} h(t - mT_c) h(t - d - mT_c) \tag{2.9}$$

In the simplest case, $E\{\alpha_m^2\} = 1$ for all values of m, however, insofar as this report is concerned, there are other cases of interest, for which corresponding results are given below. That is,

$$\alpha_m = \begin{cases} \pm 3 & \text{w.p. } 1/4 \\ \pm 1 & \text{w.p. } 3/4 \end{cases} \quad \Rightarrow \quad E\{\alpha_m^2\} = 3 \tag{2.10}$$

$$\alpha_m = \begin{cases} \pm 6 & \text{w.p. } 1/4 \\ \pm 4 & \text{w.p. } 1/2 \\ \pm 2 & \text{w.p. } 1/4 \end{cases} \quad \Rightarrow \quad E\{\alpha_m^2\} = 18 \tag{2.11}$$

$$\alpha_m = \begin{cases} \pm 7 & \text{w.p. } 1/4 \\ \pm 5 & \text{w.p. } 1/4 \\ \pm 3 & \text{w.p. } 1/4 \\ \pm 1 & \text{w.p. } 1/4 \end{cases} \quad \Rightarrow \quad E\{\alpha_m^2\} = 21 \tag{2.12}$$

These different cases only scale the basic result on the spectral contribution of the periodic component in Eq. 2.7. A complete analytical derivation on the spectrum of the output of the D & M receiver (carrier and baseband case) as a function of the actual statistics of the components of $\{\alpha_m\}$ is beyond the scope of this report. However, simulation results to be presented subsequently will demonstrate the effect (after proper scaling in order to

normalize the power of all signals) of these different statistical cases. Continuing now with Eq. 2.9, express

$$E\{\Psi(t)\} = \sum_{m=-\infty}^{\infty} E\{\alpha_m^2\} \int_{-\infty}^{\infty} H(\eta) e^{j2\pi(t-mT_c)} d\eta \int_{-\infty}^{\infty} H(\nu) e^{j2\pi(t-d-mT_c)} d\nu \qquad (2.13)$$

so that the spectral description of this averaged term is obtained from

$$\int_{-\infty}^{\infty} E\{\Psi(t)\} e^{-j2\pi ft} dt = \sum_{m=-\infty}^{\infty} E\{\alpha_m^2\} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(\eta) H(\nu) e^{-j2\pi((\eta+\nu)mT_c+\nu d)} d\eta d\nu \int_{-\infty}^{\infty} e^{j2\pi(\eta+\nu-f)t} dt \qquad (2.14)$$

Since the last integral in Eq. 2.14 is equivalent to the Dirac delta function $\delta(\eta+\nu-f)$, a significant simplification results, namely

$$\int_{-\infty}^{\infty} E\{\Psi(t)\} e^{-j2\pi ft} dt = \sum_{m=-\infty}^{\infty} E\{\alpha_m^2\} e^{-j2\pi fmT_c} \int_{-\infty}^{\infty} H(f-\nu) H(\nu) e^{-j2\pi\nu d} d\nu \qquad (2.15)$$

Clearly the expectation in Eq. 2.15 as well as the integral term are independent of the index m, and furthermore from [16],

$$\sum_{m=-\infty}^{\infty} e^{-j2\pi fmT_c} = \frac{1}{T_c} \sum_{n=-\infty}^{\infty} \delta(f - \frac{n}{T_c}) \qquad (2.16)$$

Therefore it can be seen that it is clear that Eq. 2.15 demonstrates the presence of spectral lines located at integer multiples of the chip rate frequency and strength which is affected by the receiver delay d and the HBE pulse h(t) selected. It is well known that when h(t) is a simple rectangular pulse of duration T_c, the strength of the spectral component at the chip rate frequency is maximized by setting d equal to half the duration of one chip, [5, App. G]. This optimal setting for d is however not always equal to $T_c/2$. Some studies have demonstrated that in certain circumstances, the optimum delay setting d can be equal to zero, or depending on the specific case, some value other than $T_c/2$, [17]. In fact, consider the case in which h(t) is the most bandwidth efficient pulse possible, namely

$$h(t) = \frac{\sin 2\pi t/T_c}{2\pi t/T_c} \quad \Leftrightarrow \quad \frac{T_c}{2} \text{rect}(\frac{f}{2T_c}) \qquad (2.17)$$

A number of mathematical manipulations are required to show that when the HBE pulse is given by Eq. 2.17,

$$\int_{-\infty}^{\infty} H(f-\nu)H(\nu)e^{-j2\pi\nu d}d\nu = \begin{cases} 0 & |f| > \frac{2}{T_c} \\ T_c e^{-j\pi f d}\dfrac{\sin 2\pi(\frac{1}{T_c}+\frac{f}{2})d}{\pi d/T_c} & -\frac{2}{T_c} < f < 0 \\ T_c e^{-j\pi f d}\dfrac{\sin 2\pi(\frac{1}{T_c}-\frac{f}{2})d}{\pi d/T_c} & 0 < f < \frac{2}{T_c} \end{cases} \qquad (2.18)$$

Therefore, for this particular HBE pulse, from Eqs. 2.15, 2.16, and 2.18, for $|f| < 2/T_c$, obtain

$$\int_{-\infty}^{\infty} E\{\Psi(t)\}e^{-j2\pi ft}dt = E\{\alpha_0\}\sum_{n=-\infty}^{\infty}\delta(f_c - n)e^{-j\pi|f_c|d_c}\frac{\sin 2\pi(1-|f_c|)d_c}{\pi d_c} \qquad (2.19)$$

where

$f_c = fT_c$ (normalized frequency); $d_c = d/T_c$ (delay normalized to chip duration) (2.20)

A plot of the sinc function term in Eq, 2.19 has been shown in Fig. 4, as a function of normalized frequency for values of normalized delay set to 0.1, 0.3, 0.5, 0.7, and 0.9. Observe that at a normalized frequency of 0.5, the optimum setting for the receiver delay is zero making it a simple squarer (or magnitude squared) type device. Clearly setting the normalized delay to 0.5 is not the worst choice, however since receivers would normally not have knowledge of the HBE pulse shape used by the transmitter, a compromise must be made that leads to robust performance. Therefore the choice of normalized delay set to 0.5 (which is optimum for rectangular pulse shapes) is often made and adopted here in the computer simulations carried out. It must be further understood that receivers typically have no precise knowledge about the actual value of T_c, making proper delay setting difficult at best. If some prior knowledge of the transmitter chip rate is available, the D & M device would be followed by a narrowband filter centered at the chip rate frequency whose output is applied to a threshold test for signal detection purposes. Such a device is used for declaring the presence or absence of the signal of interest. The optimal processor for signal detection and its performance in the presence of noise and interference is a separate problem beyond the scope of project.

Section 3: The Simulation System

The simulation system was built by creating individual elements that were redefined as new blocks (but called symbols in SPW) and interconnecting blocks in accordance with the system model under study. Fig. 1 shows the system at its highest hierarchical level

from a simulation standpoint. The focus was on baseband signal processing, hence no carrier modulation and demodulation blocks are present. It can be observed that the overall simulation system is made up first of random data and PN sequence generator blocks that are connected in such a way to generate the flip-wave spread signal. a random data generator. A one input/two output flip-wave symbol (labeled flpwvgen in Fig. 1) was created in order to implement the transformation specified by Eq. 1.10. The details of symbol in terms of its block construction so as to include the initialization specified by Eq. 1.11 is shown in Fig. 2. As can be seen in Fig. 1, inclusion of pulse shaping operation as well as delay and complex multiply operation is present as required in order to be able to simulate the system of interest. In order to generate both the windowed and non-windowed sinc function shaped HBE pulse of interest, finite impulse response (FIR) filters of appropriate length were used as shown (for a typical filter) in Fig. 3. Each filter block is limited in the number of coefficients that can be specified, hence for the longer duration responses parallel filters with appropriate time delay were used as shown. The system implemented used a random data generator producing a bit stream at a rate of 2 KBPS, for a bit duration T_b of 500 μsec. The output of the flip-wave generator produced the (pulse stretched) signals $X'(t)$ and $Y'(t)$ mathematically described by Eq. 1.14. These signals are mixed with separate spreading codes described by Eq. 1.17 where the chip rate is set at 256 KCPS. This means that there are 128 chips contained in each bit, or equivalently, the chip duration T_c is 3.90625 μsec. (A simulation sampling rate 20 times above the chip rate was set, that is 5.12 MSPS.) In order to conveniently generate spreading codes with amplitudes and distributions other than binary and equally likely (as specified by Eqs. 2.10, 2.11, 2.12), the outputs of three independent PN sequence generators, each having a unique LFSR length and driven by a different polynomial were weighted and summed so as to produce the desired amplitude statistics. The resulting spread signal could therefore be of binary type, or have modulation of 4, 6, or 8 amplitudes. The resulting spread signal is specified in complex form as given by Eq. 1.18 and then processed by a switch controllable carrier or baseband D & M receiver with the delay d always set to $T_c/2$. The receiver output was analyzed in terms of its spectrum via an FFT operation and the results measured directly from the spectral plots as provided by the signal analyzer feature of the simulation software. Of particular interest is the relative strength of spectral lines produced in this output at the chip rate (or in certain cases at twice the chip rate) frequency. Due to space constraints, these spectral plots are not shown here, however the measured values at specific frequencies have been tabulated in the next section.

Section 4: Simulation Results

Results of the most important simulation runs are presented in order to demonstrate the performance that can be expected of the carrier and baseband D & M receiver under various operating scenarios. The focus here is on demonstrating the detectability of the signal of interest by measuring the strength of spectral lines produced at the output of the D & M receiver. The numerical results of spectral amplitude at a given frequency are

presented in tabular form. Simulation runs were carried out in which the carrier and baseband D & M receiver operated under eight different input conditions. Under each of these eight different input conditions, four scenarios were tested in which either there was no amplitude modulation (AM) of the chips, or three different types of AM were imposed as specified by Eqs. 2.10, 2.11, and 2.12. The eight different input conditions involved selection of HBE pulse shapes, namely, a) rectangular shaped chips, b) raised cosine shaped chips, c) sinc function shaped chips of duration T_c, d) Hanning weighted sinc function shaped chips of duration T_c, e) sinc function shaped chips of duration $2T_c$, f) Hanning weighted sinc function shaped chips of duration $2T_c$, g) sinc function shaped chips of duration $4T_c$, and h) Hanning weighted sinc function shaped chips of duration $4T_c$. As a result of this, for each of the cases a) through h) considered, four sets of results were obtained for each output of the baseband D & M receiver as well as the carrier D & M receiver. The spectrum of the receiver output was obtained using the FFT evaluation feature of the simulation software and presented as a signal analysis page, (SAP). Of particular interest were discernible spectral lines (or noticeable spectral strength) in the vicinity of either the chip rate or twice the chip rate frequency, that would identify the presence of the spread signal. Therefore, spectral strength and/or spectral line presence at the frequencies of interest become the numerical parameters that yield quantitative results that become the focus of this simulation effort. The tables below present these numerical results that are further summarized and interpreted. First, it can be seen in general that the introduction of amplitude modulation to the spreading code chips tend to make the signal less detectable as can be seen from the numerical spectral strength results presented. (Due to normalization introduced to account for signal amplitude and duration changes, the tabular results can be compared against each other but not in an absolute sense due to the computational method of spectral strength.) In most of the cases considered, the baseband receiver, namely the one that delays and complex conjugates the input signal before multiplication with the undelayed version is not as effective as the carrier receiver (that does not implement the complex conjugation operation). That is, the strength of the spectral components of interest is lower at the output of the baseband versus the carrier receiver. Depending on the perspective taken, this can be a positive result. Results have shown quite clearly that longer duration sinc function shaped chips mixed with the data signal for spread spectrum modulation produce a signal with features that are more difficult to extract by D & M receivers. The Hanning window weighting not seem to affect results significantly from an LPI point of view. Moreover, raised cosine shaped chips are not very useful as studies for simple squarer intercept receivers have shown.

Table I:	Rectang. Chips Baseband Rcvr.	Rectang. Chips Carrier Rcvr.	Rsd. Cos Chips Baseband Rcvr.	Rsd. Cos Chips Carrier Rcvr.
No AM	-33.87 dB	-6.40 dB	-15.44 dB, $2R_c$	-14.32 dB
4 Level AM	-36.15 dB	-16.01 dB	-25.24 dB, $2R_c$	-23.67 dB
6 Level AM	-53.29 dB	-31.73 dB	-40.48 dB, $2R_c$	-39.72 dB
8 Level AM	-53.59 dB	-33.24 dB	-42.07 dB, $2R_c$	-41.37 dB

Table II:	Sinc T_c Chips Baseband Rcvr.	Sinc T_c Chips Carrier Rcvr.	Hanning Wght. Sinc T_c Chips Baseband Rcvr.	Hanning Wght. Sinc T_c Chips Carrier Rcvr.
No AM	-14.42 dB, $2R_c$	-10.58 dB	-19.97 dB	-19.43 dB
4 Level AM	-23.95 dB, $2R_c$	-22.41 dB	-26.49 dB	-31.34 dB
6 Level AM	-39.44 dB, $2R_c$	-35.87 dB	-43.15 dB	-46.17 dB
8 Level AM	-40.83 dB, $2R_c$	-37.40 dB	-43.18 dB	-49.96 dB

Table III:	Sinc $2T_c$ Chips Baseband Rcvr.	Sinc $2T_c$ Chips Carrier Rcvr.	Hanning Wght. Sinc $2T_c$ Chips Baseband Rcvr.	Hanning Wght. Sinc $2T_c$ Chips Carrier Rcvr.
No AM	-24.03 dB	-6.13 dB	-29.12 dB	-19.54 dB
4 Level AM	-27.69 dB	-15.53 dB	-28.69 dB	-29.03 dB
6 Level AM	-47.83 dB	-31.41 dB	-46.09 dB	-45.35 dB
8 Level AM	-52.81 dB	-32.93 dB	-46.54 dB	-46.93 dB

Table IV:	Sinc $4T_c$ Chips Baseband Rcvr.	Sinc $4T_c$ Chips Carrier Rcvr.	Hanning Wght. Sinc $4T_c$ Chips Baseband Rcvr.	Hanning Wght. Sinc $4T_c$ Chips Carrier Rcvr.
No AM	-31.88 dB	-6.98 dB	-20.47 dB, $2R_c$	-7.63 dB
4 Level AM	-37.80 dB	-16.21 dB	-29.54 dB, $2R_c$	-16.98 dB
6 Level AM	-50.46 dB	-32.81 dB	-45.26 dB, $2R_c$	-32.89 dB
8 Level AM	-66.87 dB	-33.65 dB	-46.51 dB, $2R_c$	-34.39 dB

In summary, it can be seen that techniques are available to produce signals that are nearly featureless and therefore less likely to be detected by simple D & M receivers. Due to space limitations, the actual spectral plots have not been included here but are available from the author so that more than just a single value is available on which to make performance judgments under the various operating conditions considered in this study.

CONCLUSIONS AND RECOMMENDATIONS

The research effort reported on, focuses on the performance results obtained via block oriented computer simulation of a delay and (complex) multiply receiver processing a spread spectrum signal having a specific structure. The simulations results clearly show the differences in receiver performance as the signal characteristics are changed.

The simulation results obtained are consistent with the theory and at the same time allow the investigation of system performance under different operational scenarios that could not be easily evaluated via strictly analytical means. Because of the constraints associated with this research project, the following recommendations are made regarding further investigations in this area.

1. To continue the computer modeling/simulation effort in order to evaluate the performance of receivers operating in the presence of noise and other sources of interference.
2. To evaluate the potential usefulness of other receivers types that more fully exploit the periodic statistics of the waveform under consideration, and as a result of that might exhibit superior performance or increased ability to extract signal features.
3. To study alternate methods of transmitting digital information in a spread spectrum mode that create a nearly featureless signal while maintaining sufficient signal structure so as to make a cooperating receiver practical to implement and operate.
4. To complete a detailed study followed by computer simulations quantifying receiver performance when propagation phenomena such as multipath interference are considered.

It is further proposed that a simple test bed be built that produces the signals of interest and implements the delay and multiply receiver in hardware or combination of hardware and software. Such a test bed would be field tested allowing the gathering of data against which simulation results could be compared. Field testing often reveals practical limitations that simulation efforts cannot uncover. Due to the relative ease with which the transmitter and receiver can be built, such an experimental effort is highly recommended.

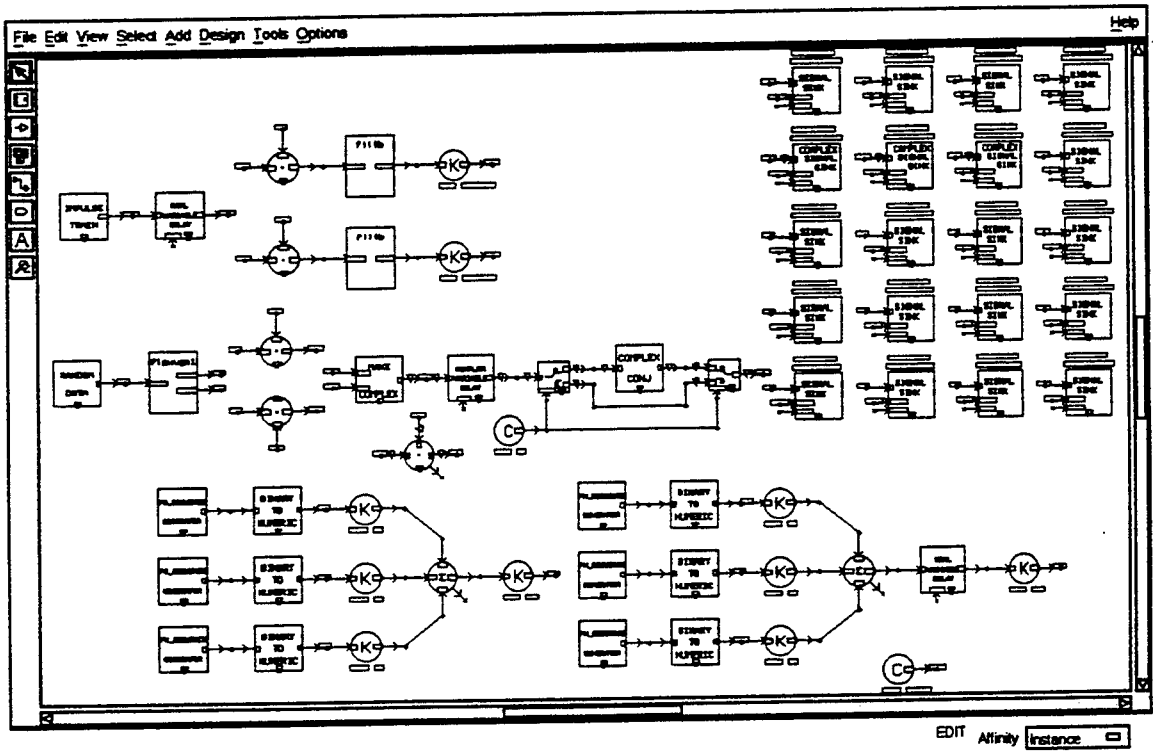


FIGURE 1: Block Diagram Design (BDD): Transmitter and Receiver System

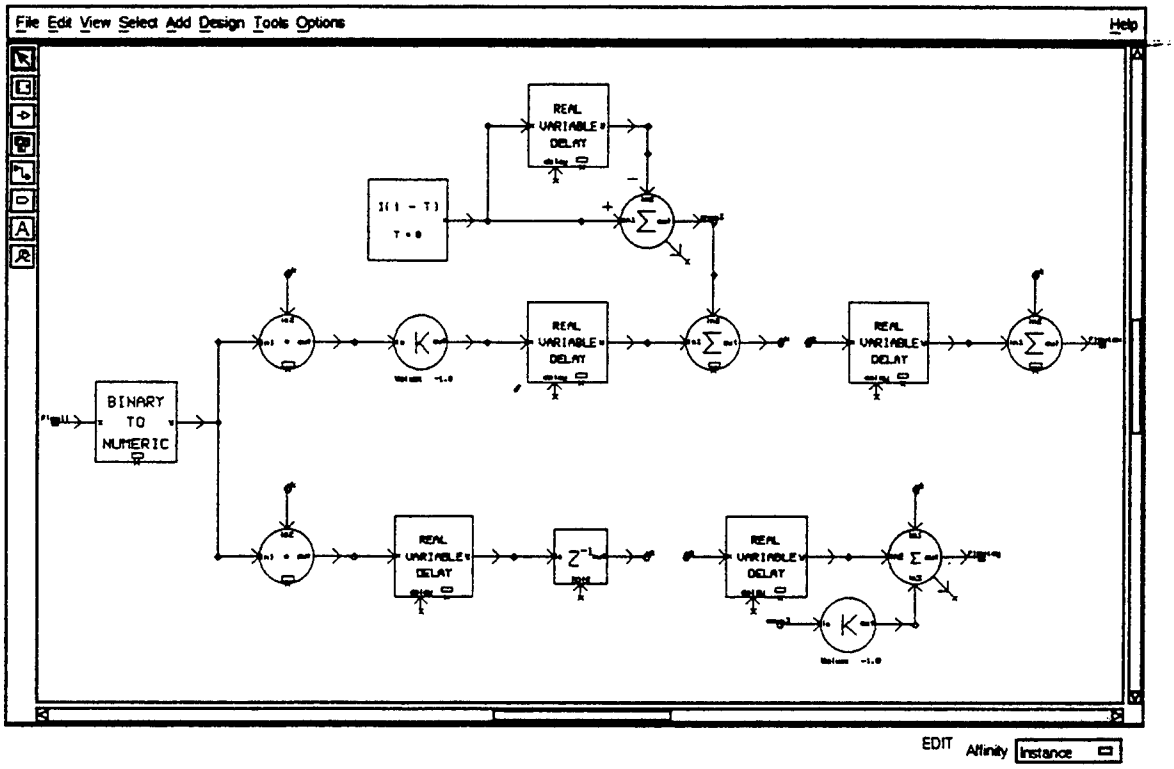


FIGURE 2: Block Diagram Design (BDD): Flip Wave Generator System

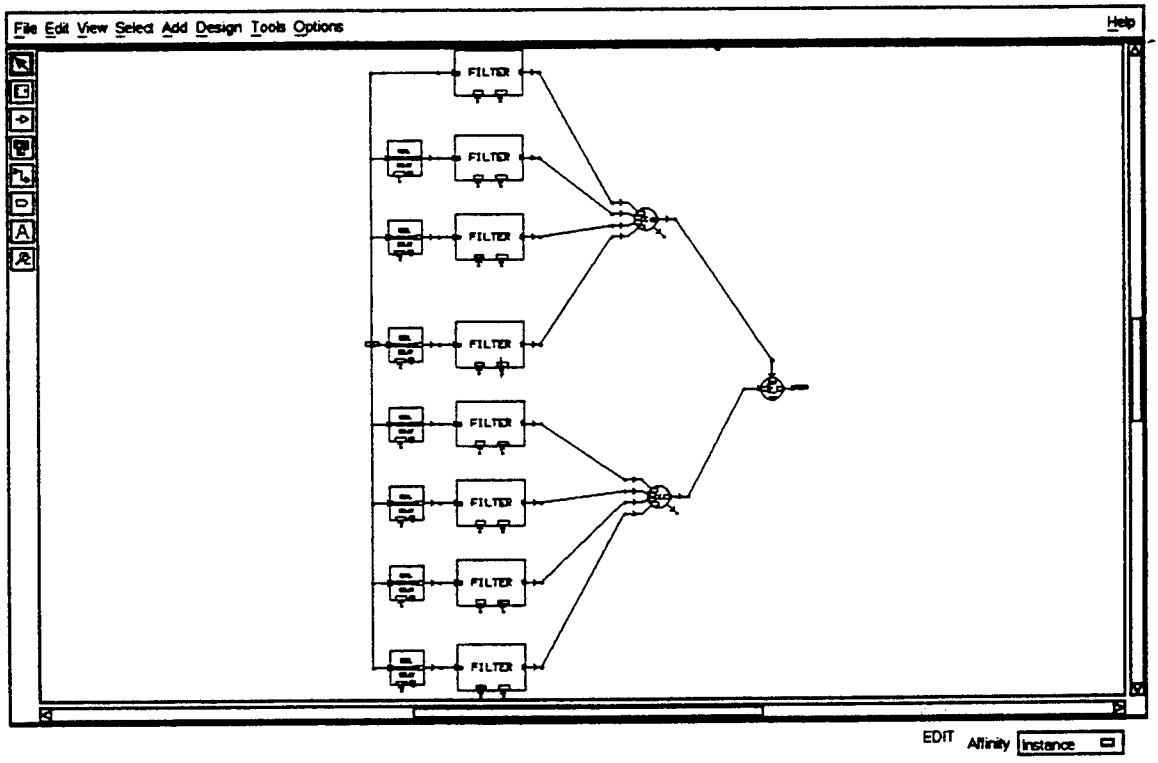


FIGURE 3: Block Diagram Design (BDD): HBE Pulse Shaping via FIR Filter

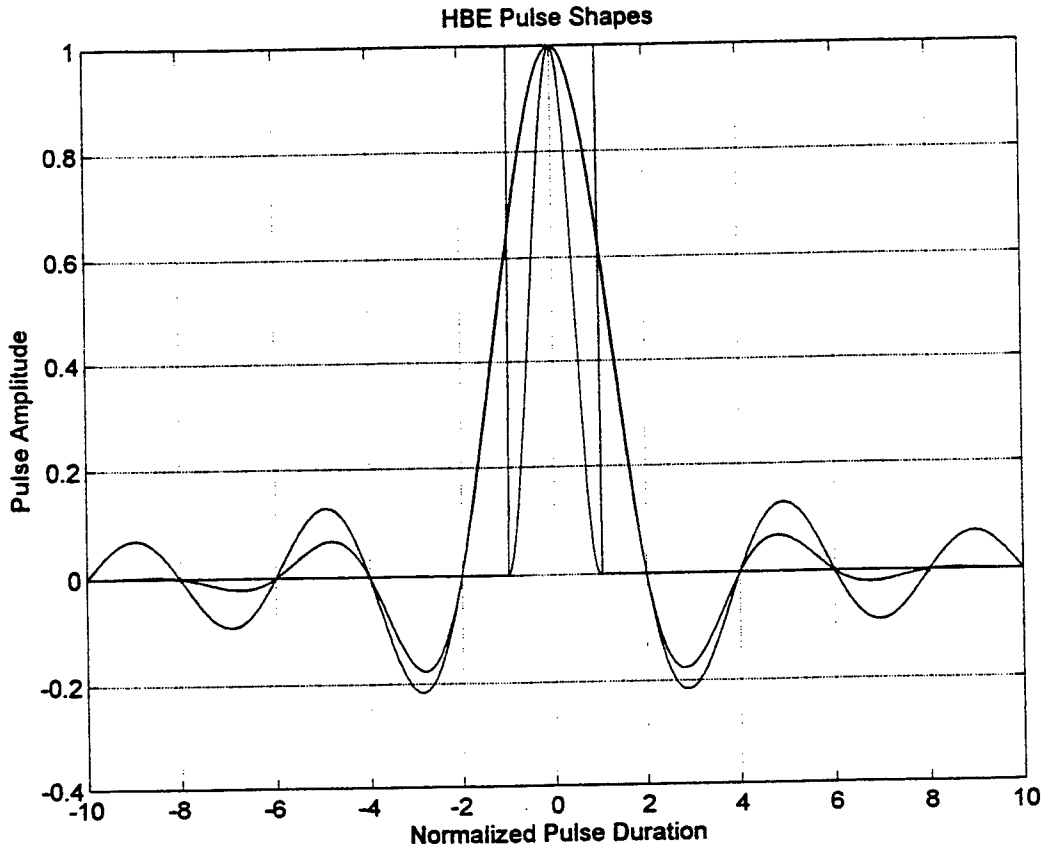


FIGURE 4: Block Diagram Design (BDD): Spectral Strength Sinc Weighting

REFERENCES

[1] deRosa, L. A., and Rogoff, M., Sec. I (Communications) of *Application of Statistical Methods to Secrecy Communication Systems*, Proposal 946, Fed. Telecommunications Lab., Nutley, N.J, Aug. 28, 1950.

[2] Sklar, Bernard, *Digital Communications*, Prentice Hall, Englewood Cliffs, N.J., 1988, Chp. 10.

[3] Cooper, G. R., and McGillem, C. D., *Modern Communications and Spread Spectrum*, McGraw-Hill, New York, N.Y., 1986, Chp. 8.

[4] Dixon, Robert, *Spread Spectrum Systems*, 3d. Edition, John Wiley & Sons, New York, N.Y., 1994.

[5] Ziemer, Rodger E., and Peterson, Roger L., *Digital Communications and Spread Spectrum Systems*, Macmillan Publishing Company, New York, 1985, Chp. 7.

[6] Gill, W. H., and Spilker, J. J., *An Interesting Decomposition Property for the Self-Products of Random or Pseudorandom Binary Sequences*, IEEE Trans. Comm., Vol. COM-11, June 1963, pp. 246-247.

[7] Kuehls, J. F., and Geraniotis, E., *Presence Detection of Binary-Phase Shift-Keyed and Direct-Sequence Spread-Spectrum Signals Using a Prefilter Delay-and-Multiply Device*, IEEE Journ. on Selected Areas in Comm., Vol. 8, No. 5, June 1990, pp. 915-933.

[8] Mauck, K. D., and Betz, J. W., *Quadratic Extraction of Features from Direct Sequence Signals*, Mitre Tech. Report, MTR 11205, October 1991.

[9] Gardner, W.A., (Editor), *Cyclostationarity in Communications and Signal Processing*, IEEE Press, Piscataway, N.J., 1994.

[10] Betz, John, *Optimal Quadratic Extraction of Cyclostationary Features*, Mitre Technical Report, MTR 11294, January 1992.

[11] Nicholson, David, *Spread Spectrum Signal Design: LPE and AJ Systems*, Computer Science Press, Rockville, MD, 1988.

[12] Reed, D. E., and Wickert, M. A., *Spread Spectrum Signals with Low Probability of Chip Rate Detection*, IEEE Journ. on Selected Areas in Comm., Vol. 7, No. 4, May 1989, pp. 595-601.

[13] Reed D. E., and Wickert, M. A., *Minimization of Detection of Baud Rate Spectral Lines by Delay and Multiply Receivers*, Proc. MILCOM 86, Monterey, CA, Oct. 1986, pp. 10.3.1-10.3.5.

[14] Patti, John, *Flip Wave -- High Bandwidth Efficiency Spread Spectrum Continuous Phase Shift Modulation System*, Technical Disclosure, Oct. 5, 1993.

[15] Patti, J., and Tyler, S., *Flip Wave Concept Demonstration*, Proc. MILCOM 95, San Diego, CA, Nov. 1995.

[16] Franks, L. E., *Signal Theory*, Prentice Hall, Englewood Cliffs, N.J., 1969, Chapter 4.

[17] Bukofzer, D. C., *Performance Analysis and System Simulations of Frequency Domain Delay and Multiply Receivers for Spread Spectrum Modulated Signals*, Final Report for Contract DAAL03-91-C-0034, U.S. Army Research Office, January 26, 1995.

OPTICAL AND NON-DESTRUCTIVE METHODS TO DETERMINE THE COMPOSITION AND THICKNESS OF AN $IN_XGA_{1-X}AS/INP$ MULTILAYER STACK

Xuesheng Chen
Assistant Professor
Department of Physics and Astronomy
Wheaton College
Norton, MA 02766

Final Report for:
Summer Faculty Research Program
Rome Laboratory/ERX

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC
and
Rome Laboratory

August 1996

OPTICAL AND NON-DESTRUCTIVE METHODS TO DETERMINE THE COMPOSITION AND THICKNESS OF AN $IN_XGA_{1-X}AS/INP$ MULTILAYER STACK

Xuesheng Chen
Assistant Professor
Department of Physics and Astronomy
Wheaton College, Norton, MA

Abstract

Two optical methods which are easy, fast and nondestructive to determine the composition and thickness of each epitaxial layer in a $In_xGa_{1-x}As$ /InP multilayer stack on InP are reported here. One is the optical reflectivity method and the other is the photoluminescence method. It is shown that the first method is very convenient and accurate as long as the model for the dependence of the refractive index of $In_xGa_{1-x}As$ on the composition x and wavelength λ is accurate. For the photoluminescence method, only preliminary result is shown here and more work needs to be done.

OPTICAL AND NON-DESTRUCTIVE METHODS TO DETERMINE THE COMPOSITION AND THICKNESS OF AN $IN_XGA_{1-X}AS/INP$ MULTILAYER STACK

Xuesheng Chen

I. Introduction

Due to its superior electronic properties, the ternary semiconductor $In_{0.53}Ga_{0.47}As$, which is lattice matched to InP, has found wide applications in high-speed electronic and optical devices such as p-i-n detectors, avalanche photodiodes and long wavelength diode lasers. Epitaxial layers of $In_{1-x}Ga_xAs$ on InP are the building blocks in these devices. The epitaxial layer composition determines both the band gap and the degree of lattice mismatch with respect to the substrate. The thickness, composition and their uniformity of the multilayer structure determine the device yield. A simple, fast, reliable and nondestructive method to determine precisely the composition, thickness and their uniformity of each layer in a multilayer stack would be very useful. The optical reflectivity method developed by Weyburne and collabrators at Rome Laboratory-Hanscom AFB has been shown recently to have this kind of desirable property for AlAs/GaAs and $Al_{1-x}Ga_xAs$ /GaAs multilayer systems[1,2]. I joined Dr. Weyburne group to extend this method to $In_{1-x}Ga_xAs$/InP multilayer stacks to see if the thickness, composition and uniformity of each layer can be determined precisely. It is essential with this method to know the dependence of the refractive index n on the alloy composition x and photon wavelength λ for $In_xGa_{1-x}As$. This information is not available in the literature. I have worked on obtaining a theoretical formula for $n(x,\lambda)$ and have got a result by modifying B. Jensen's model [3]. I also started working on the photoluminescence method to determine the composition in a $In_xGa_{1-x}As$ /InP multilayer stack.

II. Optical Reflectivity Method

II.1 Method and Theory

An $In_xGa_{1-x}As$ /InP multilayer stack consists of 7 pairs of InGaAs(~130nm)/InP(~160nm) epitaxial layers on 2 inch InP wafer with a half InGaAs cavity on top (see Fig.1). The theoretical reflectivity R [4] of a nonabsorbing multilayer stack is given by

$$R = \left| \frac{\eta_o - \frac{Y}{E}}{\eta_o + \frac{Y}{E}} \right|^2 , \qquad (1)$$

where

$$\begin{bmatrix} E \\ Y \end{bmatrix} = M_c \left(\prod_{j=1}^{m} M_j \right) \begin{bmatrix} 1 \\ \eta_s \end{bmatrix} ,$$

The transfer matrix M_k is described by

$$M_k = \begin{bmatrix} \cos\delta_k & \frac{i \sin\delta_k}{\eta_k} \\ i\eta_k \sin\delta_k & \cos\delta_k \end{bmatrix} ,$$

where $\delta_k=(2\pi n_k d_k)\cos\theta_k$, λ is the wavelength, d_k is the thickness of the k-layer, $\eta_k = n_k\cos\theta_k$ for TE mode or $\eta_k= n_k/\cos\theta_k$ for TM mode, θ_k is the incident angle and n_k is the refraction index of the k-layer. c denotes the $In_xGa_{1-x}As$ cavity, o denotes air, s for the InP substrate, and m=7 for the sample used. As you can see that the reflectivity R depends on both refractive index n_k and thickness d_k of each layer. The index of refraction formula for InP is given in Ref.[5].

It is essential to have a theoretical model to describe the dependence of the refractive index n on alloy composition x and the wavelength λ for $In_xGa_{1-x}As$

in order to use eq.2. The $n(x,\lambda)$ for $In_xGa_{1-x}As$ nearly matched to InP substrate can be obtained by using B. Jensen's model [3] with some modifications. The model uses a quantum mechanics calculations of the dielectric constant of a compound semiconductor and assumes the band structure of Kane Theory. The theoretical expressions for $n(x,\lambda)$ is given in terms of the basic material parameters of band gap energy E_g, effective electron mass m_n, effective hole mass m_p, spin orbit splitting energy Δ and lattice constant a. In the nonabsorbing range, the $n(x,\lambda)$ of $In_xGa_{1-x}As$ can be described by

$$n^2 = 1 + 2C_0 \{ (Y_B - Y_F) - z(\tan^{-1}(Y_B/z) - \tan^{-1}(Y_F/z) \} \quad , \qquad (2)$$

where

$$Y_B = m_0(a-a_0) ,$$

$$m_0 = 2.93 \text{ Å} ,$$

$$a = (1-x)\, a_{GaAs} + x\, a_{InAs} ,$$

$$a_{GaAs} = 5.6534 \text{ Å}, \quad a_{InAs} = 6.0585 \text{ Å} ,$$

$$z = [\, 1 - (\hbar\omega/E_g) \,]^{1/2}$$

$$\omega = 2\pi (c/\lambda), \qquad (c = \text{speed of light})$$

$$E_g = 1.43 - 1.53x + 0.45 x^2 ,$$

$$C_0 = (\omega_v^2/\omega_g^2),$$

$$\omega_g = E_g/\hbar ,$$

$$\omega_v = 4\pi e^2 N_v^*/m_n,$$

$$N_v^* = N_v(m_r/m_n)^{3/2},$$

$$N_v = 8/3\pi^2 \lambda_c^3,$$

$$\lambda_c = \hbar/(E_g m_n/2)^{1/2},$$

$$1/m_r = 1/m_n + 1/m_p ,$$

$$m_n = 0.07(1-x)m_e + 0.028x m_e ,$$

$$m_p = 0.5(1-x)m_p + 0.33x m_p ,$$

$m_e = 9.10939 \times 10^{-28}$ gram,

$Y_F = 2(n_e / N_v)^{1/3}$,

$n_e = 6.5 \times 10^{16}$ cm^{-3} , (carrier concentration) .

All the formulas above for $n(x,\lambda)$ are in cgs units.

The composition x and the thickness of the $In_xGa_{1-x}As$ and InP in the 7 pairs and the thickness of the $In_xGa_{1-x}As$ cavity can be easily found by treating them as adjustable parameters to get the best fit of the eq.(2) to the experimental reflectivity curve.

II.2. Results

The experimental reflectivity curve (- - -) of the multilayer stack, which consists of 7-pair $In_xGa_{1-x}As$ /InP with a $In_xGa_{1-x}As$ cavity on top, in non-absorbing wavelength range is shown in Fig.2. The best fitting curve (—) of the Jensen model (described in Section 2) to the experimental curve is also shown in Fig. 2 From the fitting, we obtained

thickness $d_{InGaAs\ in\ pairs}$=1378Å,

thickness $d_{InP\ in\ pairs}$=1547Å,

thickness $d_{InGaAs\ cavity}$=2553Å,

composition $x_{InxGa1-xAs}$ = 0.555.

These values are close to the targeted growth values. When the same experimental curve was fitted to a model (Fig.3) developed by Adachi [5], we obtained

thickness $d_{InGaAs\ in\ pairs}$=1414Å,

thickness $d_{InP\ in\ pairs}$=1502Å,

thickness $d_{InGaAs\ cavity}$=2557Å,

composition $x_{InxGa1-xAs}$ = 0.578.

The two sets of the deduced values are different, especially in the composition, due to the two different models for the $n(x,\lambda)$ of $In_xGa_{1-x}As$. Fig. 4 shows the $n(x,\lambda)$ versus λ at different x described by the modified B. Jensen's model (eq.2) and S. Adachi's model, respectively. The differences are especially noticeable in the long wavelength region. At the present time, there is no way to decide which of the two models is correct. We have compared the n values predicted by each model with the experimental indices to judge which model is better and/or to make the model better by adjusting certain parameters in the model. However, there are not enough reliable experimental indices available in our wavelength and composition range in the literature for us to compare with any model for $n(x,\lambda)$ of the $In_xGa_{1-x}As$. We are in the process of trying to find a way to measure the $n(x,\lambda)$ or effective $n(x,\lambda)$ ourselves in order to deduce precise thickness and composition from the optical reflectivity method.

In conclusion, the reflectivity methods is non-destructive, easy and fast to deduce precisely the thickness and composition of each layer in a multilayer stack as long as the model for $n(x,\lambda)$ or effective $n(x,\lambda)$ is accurate.

II. Photoluminescence Method

It was shown in Ref.[6] that the wavelength position on the lower energy side at a half of the peak photoluminescence intensity shifts with the composition x of the $In_xGa_{1-x}As$ epitaxial layer on InP. It would be easy and nondestructive if we can use this method to determine the composition and its uniformity of each epitaxial layer in the multilayer $In_xGa_{1-x}As$ /InP stack. Preliminary work has been done in this direction. Fig.5 shows the photoluminescence from the $In_xGa_{1-x}As$ on InP, excited by He-Ne laser (632.8nm). The InGaAs layer thickness is about 1.2um. However, I was not able to observe any photoluminescence signal from the InGaAs/InP multilayer stack. One reason might be due to that the thickness of each $In_xGa_{1-x}As$ layer in the stack is too small (~0.1um) so that the signal is too

small to detect. It also could be that the experimental set up was not well arranged to detect weak signal. I realigned the set-up with InP(excited by He-Ne laser) and was able to increase its photoluminescence signal significantly. I will try again to see if I can detect any photoluminescence signal for the $In_xGa_{1-x}As$ /InP multilayer stack. I would like to continue to work on determining the composition and its uniformity of the multilayer system using photoluminescence method.

References

[1] Qing S. Paduano, David Weyburne, Fenglu and R. Bhat, J. Elect. Mat. Vol.24, No.11, 1995, p1659.

[2] Qing S. Paduano and David Weyburne, et al., private communication.

[3] B. Jensen and A. Torabi, J. Appl. Phys. Lett. Vol.33, No.7, 1978, p659.

[4] H. A. MaLeod, Thin-Film Optical Filters (Bristol: Hilger, 1986), p32.

[5] S. Adachi, J. Appl. Phys. Vol.53, No.8, 1982, p5863.

[6] I. C. Bassignane, C. J. Miner, and N. Puetz, J. Appl. Phys. Vol.65, No.11, 1 June 1989, p4299.

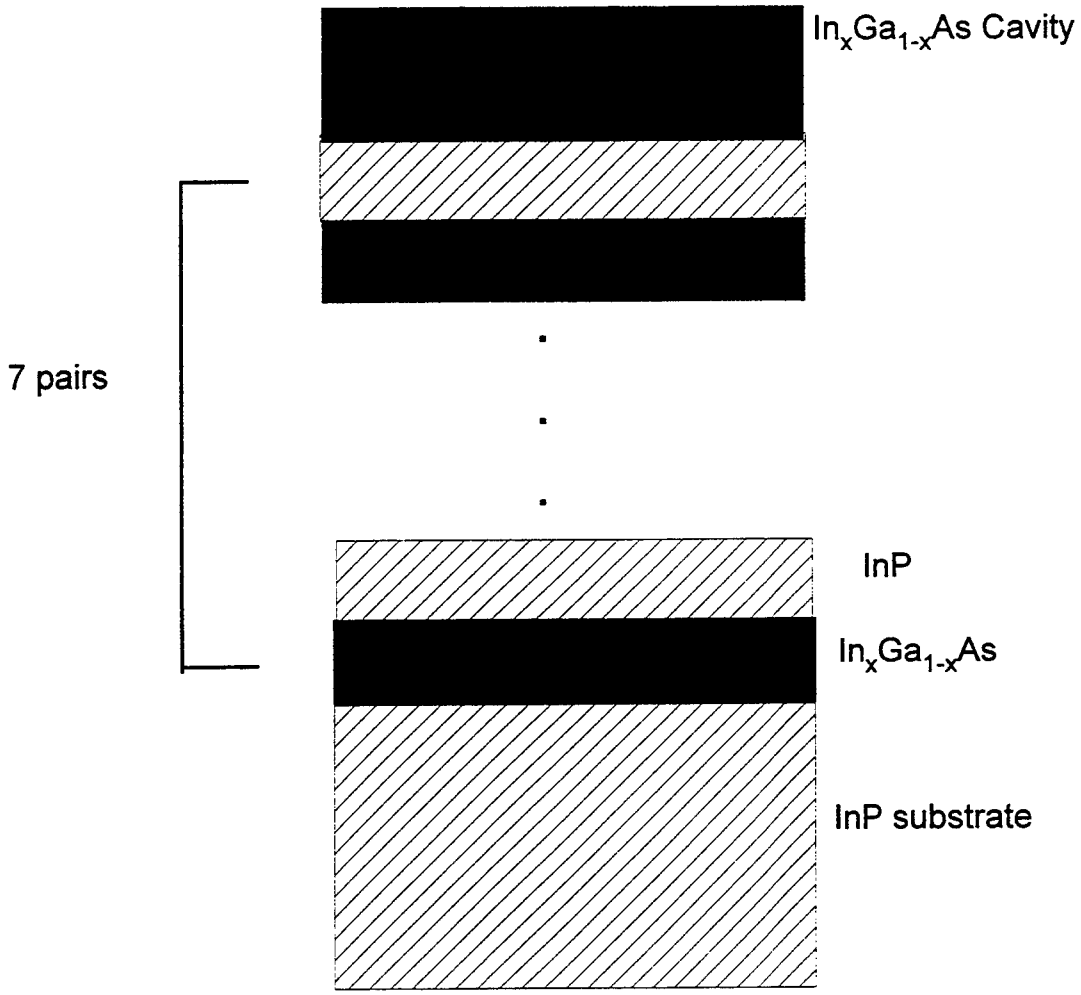


Fig. 1. The cross section of the $In_xGa_{1-x}As/InP$ multilayer stack.

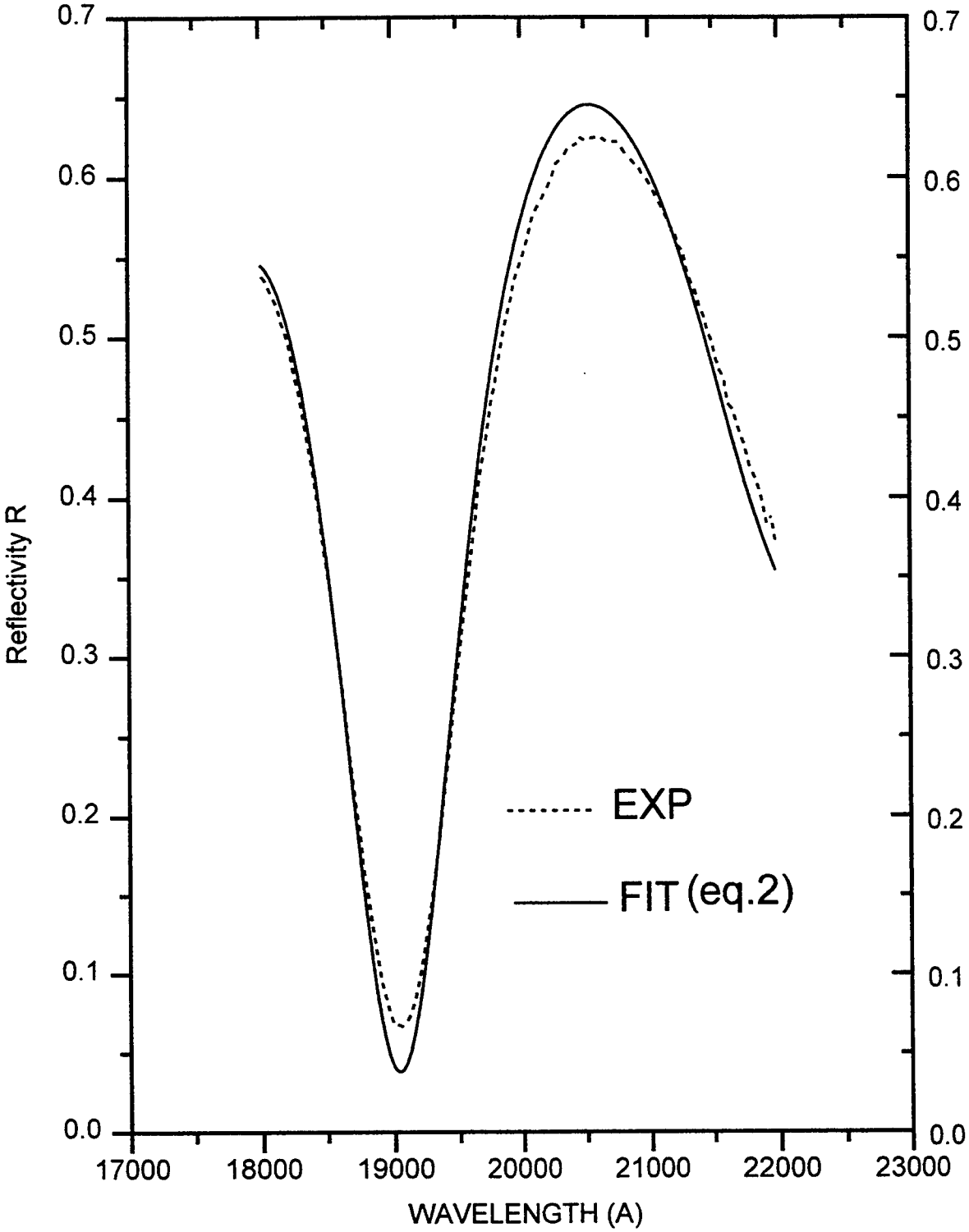


Fig.2. The experimental reflectivity curve(- - -) and the best fitting curve (solid line) using eq.2 for the InGaAs/InP multilayer.

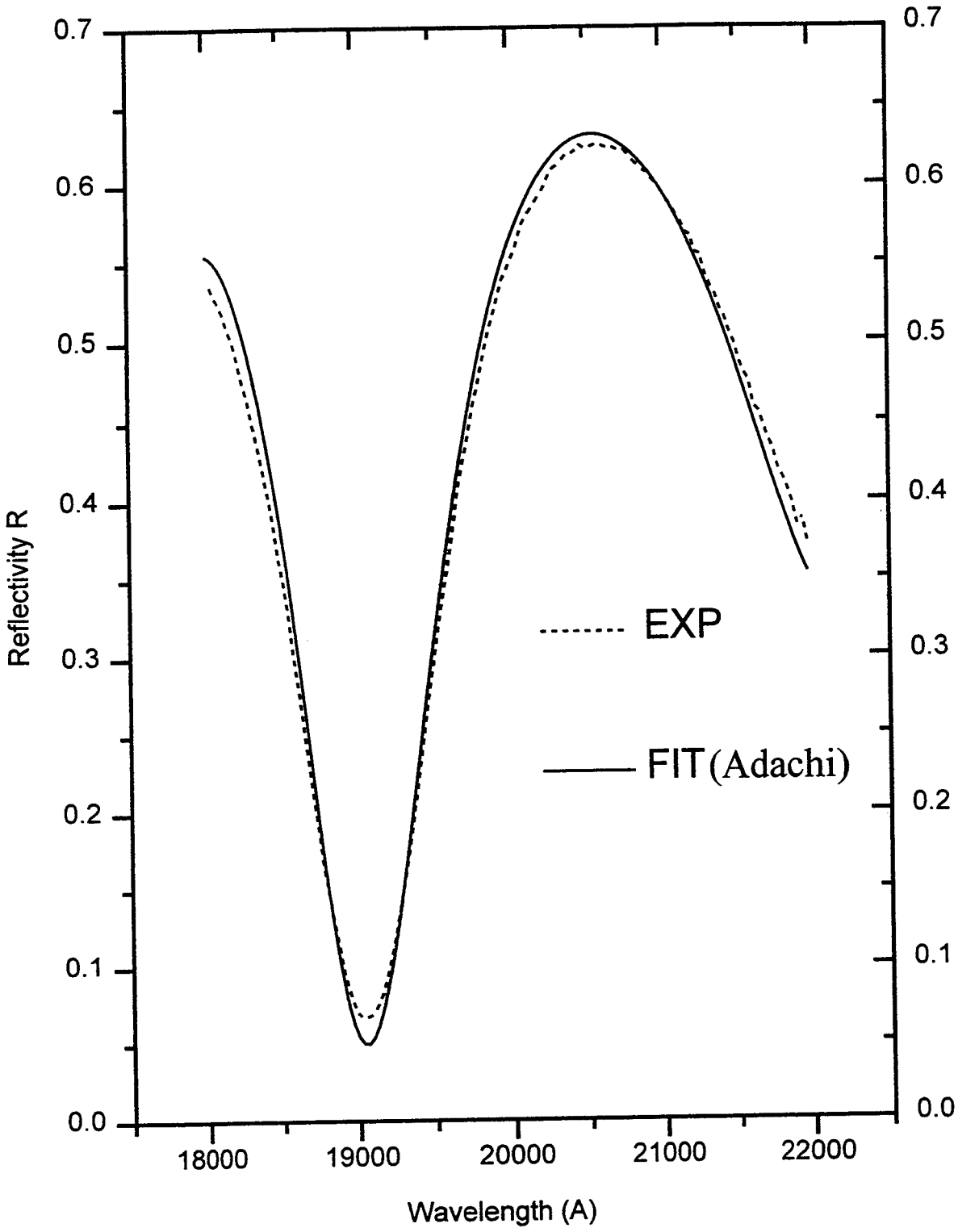


Fig.3. The experimental reflectivity curve (- - -) and the best fitting curve (solid line) using Adachi's model.

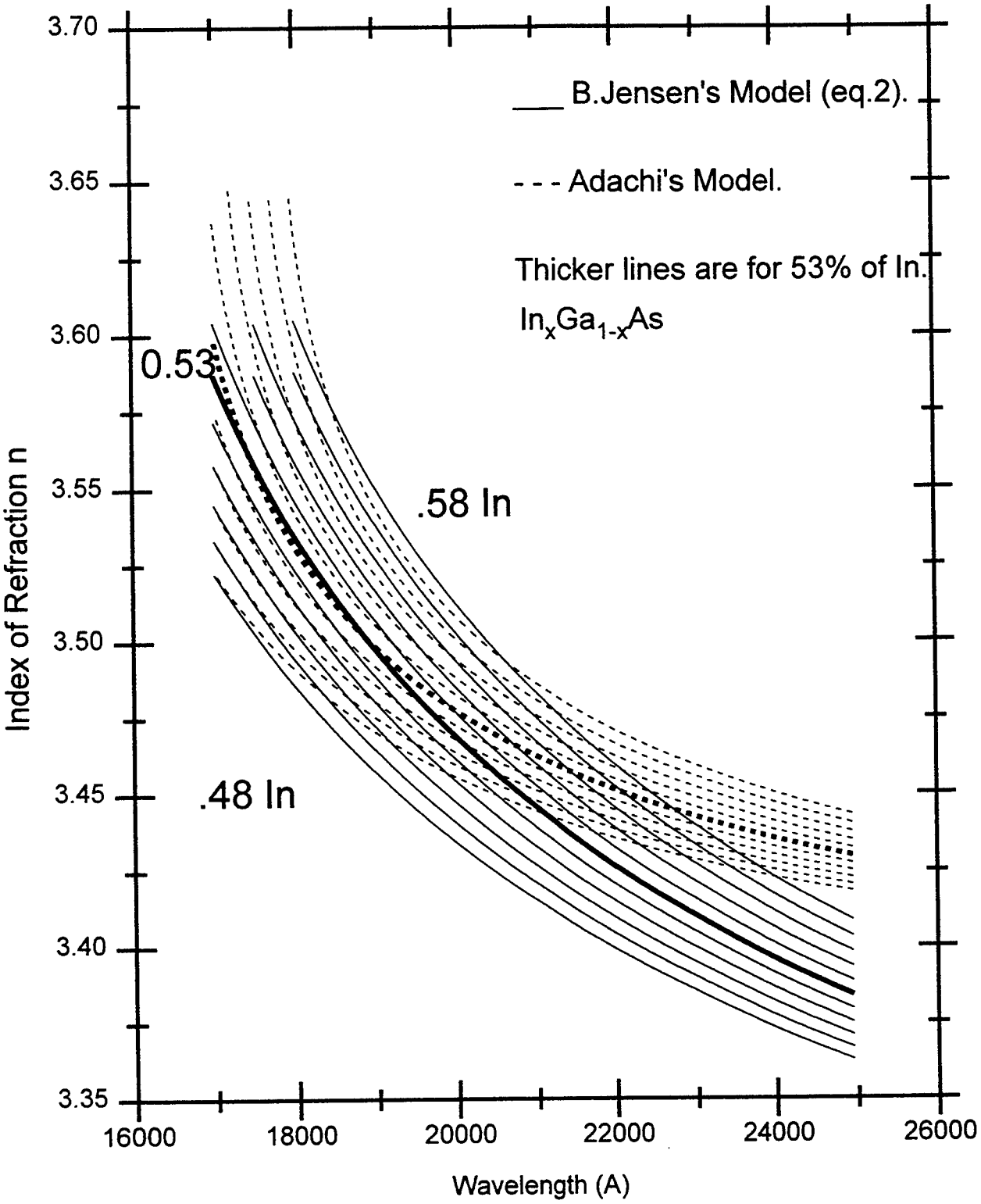


Fig.4. The index of refraction $n(x,\lambda)$ versus λ at different composition x using B.Jensen's model and Adachi's model, respectively.

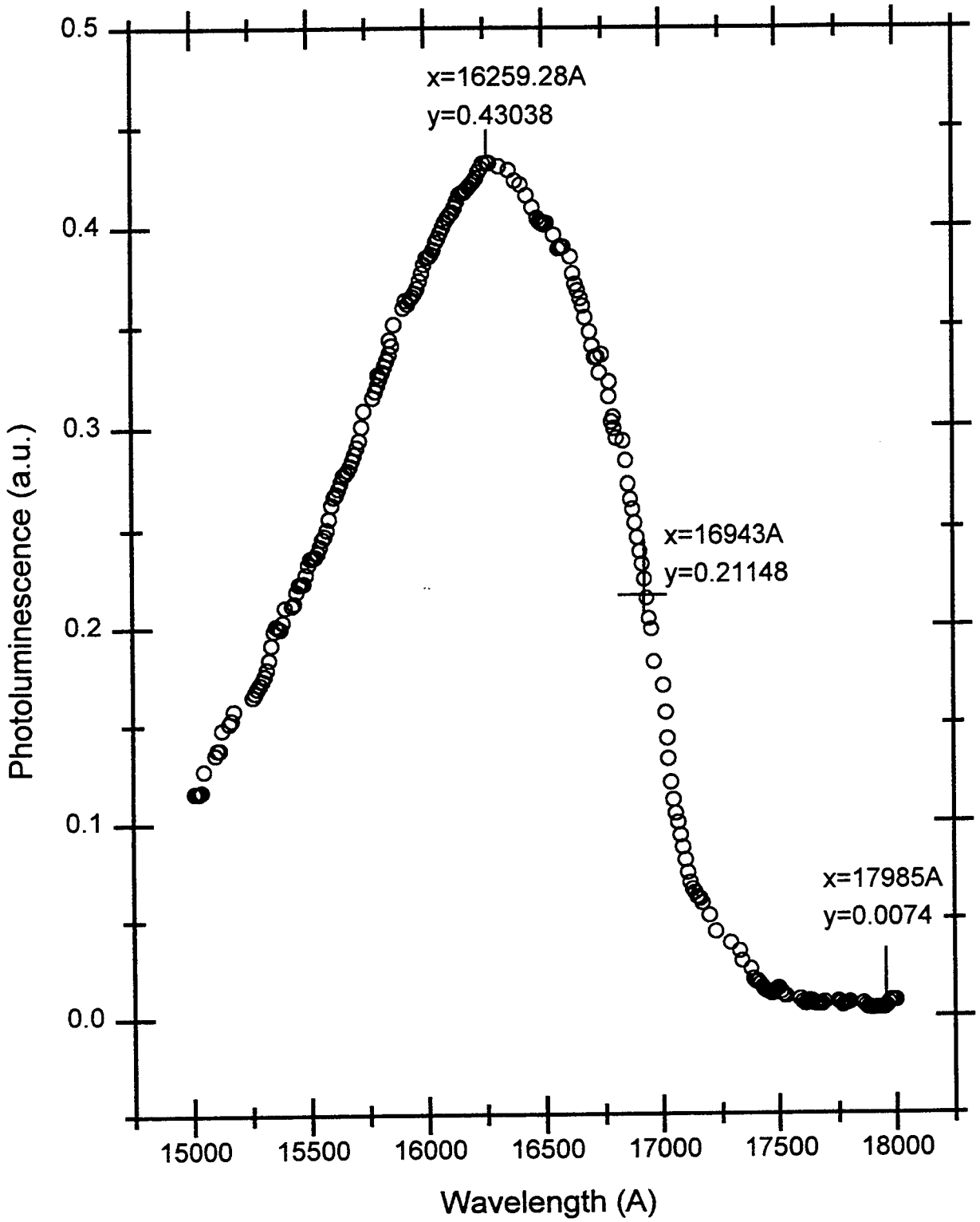


Fig.6. Room-temperature photoluminescence of a 1.2μm-thick $In_xGa_{1-x}As$ on InP.

A STUDY OF OPTOELECTRONIC FEEDBACK-SUSTAINED PULSATION OF LASER DIODES AT 1300 nm AND 780 nm

J. Chen

Research Scientist

Department of Electrical and Computer Engineering
Rochester Institute of Technology, Rochester, NY 14623

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office Of Scientific Research
Bologna Air Force Base, DC

and

Rome Laboratory, Rome, New York

September 1996

A STUDY OF OPTOELECTRONIC FEEDBACK SUSTAINED PULSATION OF LASER DIODES AT 1300 nm AND 780 nm

by

J. Chen, G. Li, F. Tian

Rochester Institude of Technology, Rochester, NY 14623

J. Kann, R. K. Boncek, D. J. Grucza

Rome Laboratory, Rome, NY 13441

Abstract

In this report, some new experimental results are presented on optoelectronic feedback sustained pulsation in multi-quantum well InGaAsP Laser diodes at 1300 nm and AlGaAs injection laser diodes at 780 nm. The feedback intensity plays an important role in feedback sustained pulsation in these two different kinds of laser diodes. It was found a bistability of the feedback oscillating modes appeared as increasing the feedback intensity. It was observed that jumps between feedback oscillating modes occurred as varying the drive current. However, in between jumps, the frequency response to the drive current is not zero, instead, it depends on the feedback intensity. In addition, amplitude modulations in LD at 1300 nm are demonstrated using FSP at 1 GHz as the subcarriers. .

1. Introduction

Recently, lots of efforts have been made for all optical SCM/SCMA networks in which optical technology including subcarrier generation plays an important role in network functionality[1,6]. The key device for optical subcarrier generation is the self-sustained pulsation(SSP) or feedback-sustained pulsation(FSP) laser diode(LD). Actually, the SSP LD is a nonlinear dynamical system if the coupling between gain and absorption is concerned[7-11]. In a nonlinear dynamical system , the creation of periodic orbits as varying a control parameter is the most important bifurcation process[12]. In the case of SSP LD, the control parameter is the drive current for LD. It determines the eignvalues of the system and when a Hopf Bifurcation occurs it determines the frequency and amplitude of the periodic orbits of the system. In other words, the frequency and amplitude of the periodic orbits can be tuned by directly varying the drive current for communication purpose. Information can be impressed on the frequency and/or amplitude of the periodic obits through the driving current. Therefore, as a current-controlled oscillator(CCO), SSP LD can generates tunable MMW subcarriers on an optical carrier without any microwave components.

Many efforts have been initiated to the synchronization and feedback of SSP LD. [13-20] since the discovery of SSP[21], because synchronization and feedback (both optical and optoelectronic) can stabilize SSP and decrease the pulse width as well as improve jitter of SSP. The wavelength division multiplexing(WDM)-subcarrier multiplexing (SCM) are of particular interest for local area networks(LAN) where pure WDM is not able to provide a large number of nodes. This motivated us to focus our attention on how to generate microwave subcarriers in laser diodes other than 780 nm. Earlier results about the transient SSP and optoelectronic FSP in InGaAsP single-mode LD at 1300 nm were published elsewhere[20]. In this report, some new experimental results of FSP in multi-quantum wells(MQW) LD at 1300 nm as well as AlGaAs injection LD at 780 nm are

reported for the first time. It was found that there exists a bistability of feedback oscillating modes as the drive current ramped up and down in LD both at 1300 and 780 nm. It was observed that as the drive current changed the oscillating mode jumped abruptly, but in between jumps, FSP frequency still increased with drive current but with a much slower rate than SSP. This frequency response to the drive current depends on the feedback intensity. Based on the flatness of the frequency response to the drive current in MQW LD at 1300 nm, a demonstration of amplitude modulation was presented also in this report.

2. Experimental set-up

The experimental set-up is illustrated schematically in Figure 1. The pigtailed fiber from LD for SSP and FSP measurements was fusion spliced to a 3-dB bi-directional fiber coupler. Half of the laser output was used for power measurements or monitoring waveform of SSP or FSP in time domain by an HP 54120A oscilloscope. The other half laser output was fed back through an optoelectronic feedback loop. In this loop, the laser beam was first introduced to a New Focus 1414B pin photodiode(PD) which converts SSP or FSP into microwave signals. The microwave signals were pre-amplified(with a 30 dB gain) and directed to a K & L 5DH1-1000/10000 1-10 GHz high pass filter(HPF). After the HPF, the signals were split into two parts. One part was used to monitor the spectrum of the signals with an HP 8593E microwave spectrum analyzer. The other part of the preamplified signals was firstly directed into an attenuator-power amplifier(with a 30 dB gain)-attenuator parallel series so as to adjust the feedback intensity, and then connected onto a power splitter. The output terminal of this splitter was connected to an Aventech AVX-SRB Bias Tee for laser diode. The other input terminal of this splitter was connected to an HP 8657B signal generator to get modulating signals. Therefore, the feedback as well as modulating signals mixed together within this splitter and finally directed to the Bias Tee to drive the laser diode. The drive current for LD was provided by an ILX LDC-3722 LD controller. The laser diodes were temperature stabilized at 22.7 °C with a thermal electric cooler. The delay time of the feedback loop can be adjusted by changing the length

of both optical fibers or co-axial microwave cable. In the case of LD at 780 nm, because its SSP is much narrower and stronger than that from MQW LD at 1300 nm, it is not necessary to use high pass filter to get strong FSP signals.

3. Experimental results

The SSP in MQW LD at 1300 nm is hardly detected by oscilloscope compared to AlGaAs injection LD at 780 nm.. However, it can be detected by spectrum analyzer. The band width of the fundamental SSP is about 500 MHz. The peak power of the fundamental component ranges from -40 to -50 dBm. The average SSP frequency change per unit current was estimated as around 180 MHz/mA, and 60 MHz/mA for MQW LD at 1300 nm , and AlGaAs injection LD at 780 nm, respectively.

By adjusting attenuation in the loop, that is, changing the feedback intensity, a very strong FSP can be obtained. The spectrum of FSP of MQW LD is almost same as that of InGaAsP single-mode LD at 1300 nm, which was described in more details in [20]. But no transient SSP was observed in MQW LD. The harmonic components of FSP are as strong as the fundamental's. This means that the pulse shape is not a sinusoidal one as revealed in the time domain. Usually, there are 5 to 7 feedback modes oscillating simultaneously. The frequency spacing between any two adjacent modes is just the inverse of the delay time of the feedback loop including the optical as well as electronic delay. However, the central or principal oscillating mode can be 40 dB stronger than the other sideband oscillating modes by adjusting the feedback intensity or driving current. Because the amplitude of periodic obits depends on the driving current, the intensity of the feedback signals will change as varying drive current even with a constant loop gain. For example, the second harmonic component can be 30 dB lower than the fundamental's at 27 mA with a loop attenuation of 5 dB in the case of MQW LD compared to 16 dB at 17 mA.

The frequency of the fundamental FSP versus drive current in the case of LD at 780 nm is shown in Figure 2. The SSP's is also shown in Figure 2 for comparison. The SSP frequency is tuned almost linearly with the driving current. But FSP frequency is almost a periodic step function of the driving current. The period of each step is just the spacing of the two adjacent feedback oscillating modes, which is the inverse of the delay time of the optoelectronic feedback loop. The average mode spacing was estimated as 140 ± 10 MHz. The jumps appears very abruptly at least within a current range of less than 0.1 mA, which is the adjusting limit of the current controller. However, in between jumps, that is, on the steps there definitely exists an FSP frequency response to the drive current, though it changes slowly and almost linearly in response to the varying current. The slope of the steps on the picture is not zero, instead it ranges between 13 to 30 MHz per mA. It was found that the FSP frequency response to the current, that is, the slope, on the steps does depend on the feedback intensity. As the loop attenuation varying from 10 dB to 6 and then to 3 dB, the slope of steps changed from 13-30 MHz/mA to 10-20 MHz/mA and then to 7-10 MHz/mA, correspondingly.

In the case of MQW LD, usually the feedback mode only oscillates at a frequency near the cut-off frequency of the high pass filter in the feedback loop. But if increasing the feedback intensity, one, even two mode jumps can be obtained as varying the drive current. The FSP frequency response to unit drive current is estimated as low as 0.13 MHz/mA, which is more than three order of magnitudes lower than SSP's. The FSP DC output power responses in very good linearity to the drive current within the maximum current limit as shown in Figure 3. The fundamental FSP power increases linearly with the drive current beyond 18 mA. A typical characteristic curve is shown in Figure 4. A very flat response of frequency and a linear response of output power to drive current makes it possible to do amplitude modulation by means of FSP as subcarriers in MQW LD. The AM will be described in some details later in this letter.

As mentioned above, increasing the feedback intensity results in a relatively flat FSP frequency response to the drive current. It also results in bistability of FSP. Figure 5a is a typical picture showing this FSP bistability in the case of AlGaAS LD at 780 nm. When increasing the drive current from 65 to 66.9 mA, the FSP keeps oscillating on the same feedback mode. But the frequency increases linearly with the drive current from 869 to 893 MHz. As the drive current furtherly increases a little bit, for example 0.1 mA, reaching 67 mA, the FSP jumps up to the adjacent feedback oscillating mode with a high frequency of 994 MHz. As the drive current ramps up again, the FSP keeps oscillating on this high frequency mode till jumps up to the next high feedback oscillating mode. As the driving current ramps down from 68 mA at 1008 MHz, the FSP keeps oscillating on this high frequency mode till jumps down to the next low frequency oscillating mode at 65.9 mA instead of 67 mA, as mentioed above, which is the jump up point. A hysteretic loop is clearly seen from Figure 5a. In the case of MQW LD at 1300 nm, same hysteretic effect was also oberved except with different frequency response to the drive current as shown in Figure 5b. It was estimated from these two pictures that the frequency response to the drive current in the case at 1300 nm(0.13 MHz/mA) is much smaller that the case at 780 nm(13 MHz/mA).

Furthermore, it was found that the feedback intensity can change the depth of this frequency hysteretic phenomena. In the case of injection LD at 780 nm, as shown in Figure 5a, the drive current limit range corresponding to the hysteretic loop was expanded to 2 mA instead of 1 mA when the feedback loop attenuation decreased from 10 dB to 3 dB.

As mentioned early, by changing the attenuation in the feedback loop, the FSP waveform changes. In the case of MQW LD, by increasing the drive current the waveform

of FSP can get improved till at one point it become almost sinusoidal. At this point, the harmonic components of FSP drops a lot . The following list indicates how the attenuation in the loop effects the waveform of FSP. In the list, I_d is the driving current. $P^{1\omega}$ is the power of the fundamental FSP, and $P^{2\omega}$ is the power of the second harmonic FSP.

Attenuation(dB)	$P^{1\omega}$ (dBm)	$P^{2\omega}$ (dBm)	I_d(mA)
3	-17.8	-46	41.3
4	-20.5	-49	33.8
5	-23.3	-47	27.3

All three second harmonic powers are almost 25 to 30 dB below the fundamentals' though the corresponding drive currents are much different by merely changing the loop attenuation a little bit.

The amplitude modulation was demonstrated by applying the modulating signal (sinusoidal, 2 MHz) together with the feedback signals to the MQW LD because of its flat frequency response to the drive current. The modulated side-band power measured with the RF spectrum analyzer is proportional to the power of the modulating signals as shown in Figure 6.

It was observed that the modulation efficiency increased with the drive current till reached a maximum, and then decreased as the drive current continuous to increase while keeping a constant modulating singal. It is shown in Figure 7. If we define the modulation efficiency as the ratio of the measured power of modulated sideband signals over the measured power of the fundamental FSP

$$M_e = P_{mod}/P_{fd} \qquad (1),$$

a maximum modulation efficiency of about 40% was obtained, which is almost the same as that obtained from the waveform measurements in the time domain.

Discussion and Summary

As mentioned previously the SSP LD is a dynamical system if the coupling between absorption and gain is concerned. In the optoelectronic FSP experiments as described above, the feedback intensity plays an important role in characterizing FSP. As a feedback current controll oscillator, in order to sustain the oscillation, the phase and the gain condition should be satisfied in the SSP LD system. Our results show that on the one hand, the feedback mode jumps are abruptly, but on the other hand, there is a non-zero mode-frequency response to the drive current and it depends on the feedback intensity. This means that the dynamic process in this system is disturbed by the feedback signals. Therefore, more theretical work need to be done to explain these experimental results. As the feedback intensity increases, some nonlinear properties like bistability occurs as described above. This implies us to consider the FSP system as another dynamical system, in which the SSP LD is a generalized current control oscillator, the feedback signals are used as self-modulation of the carrier density or SSP itself. But this is an interconnected or double dynamic system. Like in other dynamic laser systems[22 - 24], as increasing the feedback intensity, the FSP lead to chaos through the period-doubling route or other routes can be expected. The theoretical as well as experimental work in search for universality in behavior of this nonlinear system and their transitions to chaos is undergoing.

Acknowledgement

J. Chen thanks James W. Cusack and Andrew R. Pirich for their help and support during the AFOSR SRP.

Figure Caption

Figure 1. Schematic of the experimental setup for studying SSP, FSP, and AM in Lds both at 1300 nm abd 780 nm. Here, 1: Bias Tee; 2: Bi-directional Fiber Coupler; 3: Optical Power Meter or Optical Spectrum Analyzer; 4: Pin Photo-Diode; 5: Pre-amplifier; 6: High Pass Filter; 7: RF Power Splitter; 8: RF Spectrum Analyzer or Oscilloscope; 9: Attenuator; 10: Power Amplifier; 11: Adjustable Attenuator; 12: RF Power Splitter; 13: High Frequency Signal Generator; 14: LD Current Controller.

Figure 2. The frequency response of the fundamental FSP and SSP to the drive current in AlGaAs injection LD at 780 nm..

Figure 3. Average FSP DC output power of MQW LD versus drive current.

Figure 4. The power of the fundamental FSP at 1300 nm as a function of the drive current.

Figure 5. Frequency response of the fundamental FSP to drive current showing the bistability of FSP in a) AlGaAs injection LD at 780 nm, and b) MQW In GaAsP LD at 1300 nm. Here, symbols of □ and Δ represent the measurements of increasing and decreasing the drive current, respectively.

Figure 6. The first sideband power of the amplitude modulated signals versus the power of modulating signals.

Figure 7. Time domain waveform of the modulated FSP signals showing the nearly sinusoidal AM with a modulation efficiency of 35% at 2 MHz.

References

1. G. Li, F. Tian, and R. K. Boncek, "Theoretical Foundation of the Optical Current-Controlled Oscillator and Its Applications in All-Optical Subcarrier Multiplexed Networks", submitted to *J. Lightwave Technol.*, May 1996.

2. X. Wang, G. Li, and C. S. Ih, "Microwave/Millimeter-Wave Frequency Subcarrier Lightwave Modulations Based on Self-Sustained Pulsation of Laser Diodes," *IEEE J. Lightwave Technol.*, **LT-11**(2), 309(1993).

3. J. B. Georges and K. Y. Lau, "800 Mb/s Microwave FSK Using a Self-Pulsating Compact Disk Laser Diodes," IEEE Photon. Technol. Lett. **4**(6), 662(1992).

4. J. . B. Georges and K. Y. Lau, "Self-Pulsating Laser Diodes as Fast-tunable(1ns) FSK Transmitters in Subcarrier Multiple-Access Networks," *IEEE Photon Technol. Lett.*, **5**(2), 242(1993).

5. Guifang Li, "Optical Generation and Frequency Modulation of Microwave Subcarrier for Antenna Remoting and Phase Array," invited Paper *SPIE Pro.* **2481**, 62(1995).

6. P. E. Barnsley, H. J. Wickes, G. E. Wickes, and D. M. Spirit, "All-Optical Clock Recovery from 5Gb/s RZ Data Using a Self-Pulsating 1.56 μm Laser Diode," *IEEE Photon. Technol. Lett.* **3**(10), 942(1991).

7. R. W. Dixon and W. B. Joyce, "A possible Model for Sustained Oscillations (Pulsations) in (Al,Ga)As Double-Heterostructure Lasers," *IEEE J. Quantum Electron.* **QE-15** (6), 4701979).

8. M. Ueno and R. Lang, "Conditions for Self-sustained Pulsation and Bistability in Semiconductor Lasers," *J. Appl. Phys.* **58**(4), 1689(1985).

9. E. A. Avrutin, "Analysis of Spontaneous Emission and Noise in Self-Pulsing Laser Diodes," *IEE Pro. -J*, **140**(1), 16(1993).

10. M. Yamada, "A Theoretical Analysis of Self-Sustained Pulsation Phenomena in Narrow-Strip Semiconductor Lasers," *IEEE J. Quantum Electron.*, **QE-29**(5), 1330(1993).

11. G. H. M. van Tartwijk and M. S. Miguel, "Optical Feedback on Self-Pulsating Semiconductor Lasers," *IEEE J. Quantum Electron.*, **QE-32**(7), 1191-1201(1996).

12. S. H. Strogatz, Nonlinear Dynamics and Chaos. (Addison-Wesley, Rwading, MA 1994)

13. A. Egan, P. Rees, J. O'Gorman, M. Harley-Stead, G. Farrell, J. Hegarty, and P. Phelan, "Theoretical Investigations of Electro-optical Synchronization of Self-pulsating Laser Diodes," *IEE Pro. -Optoelectron.*, **143**(1), 1996.

14. Y. Simler, J. Gamelin, and S. Wang, " Pulsation Stabilization Enhancement in Self-Pulsating Laser Diodes," *IEEE Photo. Technol. Lett.*, **4**(4), 329(1992).

15. K. Y. Lau and A. Yariv, "Self-Sustained Picrosecond Pulse Generation in a GaALAs Laser at an Electrically Tunable Repetition rate by Optoelectronic Feedback," *Appl. Phys. Lett.*, **45**(2), 124(1984).

16. M. Nakazawa, M. Tokuda, and N. Uchida, "Self-Sustained Intensity Oscillation of Laser Diode Introduced by a Delayed Electrical Feedback Using an Optical Fiber and an Electrical Amplifier," *Appl. Phys. Lett*, **39**(5), 397(1981).

17. K. Y. Lau, L. Figueroa, and A. Yariv, "Generation and Quenching of Intensity Pulsations in Semiconductor Lasers Coupled to External Cavities," *IEEE J. Quantum Electron.*, **QE-16**(12), 1329(1980).

18. T. C. Damen and M. A. Duguay, "Optoelectronic Regenerative Pulser," *Electron. Lett.* **16**(5), 166(1980).

19. T. L. Paoli and J. E. Ripper, "Frequency Stabilization and Narrowing of Optical Pulses from CW GaAs Injection Lasers," *IEEE J. Quantum Electron.*, **QE-6**(6), 335(1970).

20. G. Li, R. K. Boncek, X. Wang, and D. H. Sackett, "Transient and Optoelectronic Feedback-Sustained Pulsation of Laser Diodes at 1300 nm," *IEEE Photon. Technol. Lett.* **7**(8), 854(1995).

21. T. L. Paoli and J. E. Ripper, "Coupled Longitudinal Mode pulsing in Semiconductor Lasers," *Phys. Rev. Lett.* **22**, 1085(1969).

22. J. Sacher, D. Baums, P. Panknin, W. Elsasser, and E. O. Gobel, "Intensity Instability of Semiconductor Lasers under Current Modulation, External Light Injection, and Delayed Feedback," *Phys. Rev.* **A45**, 1893(1992).

23. T. B. Simpson, J. M. Liu, A. Gavrielides, V. Kovanis, and P. M. Alsing, "Period-doubling Route to Chaos in a Semiconductor Laser Subject to Optical Injection," *Appl. Phys. Lett.* , **64**, 3539(1994).

24. T. B. Simpson, J. M. Liu, A. Gavrielides, V. Kovanis, and P. M. Alsing, "Period-doubling Cascades and Chaos in a Semiconductor Laser with Optical Injection," *Phys. Rev.*, **A51**, 4181(1995).

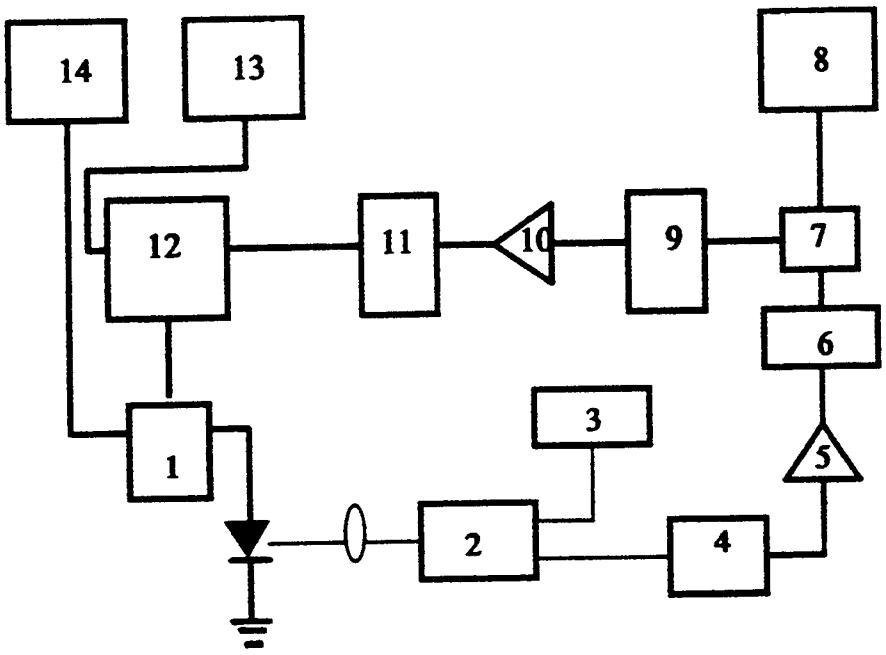


Fig. - 1

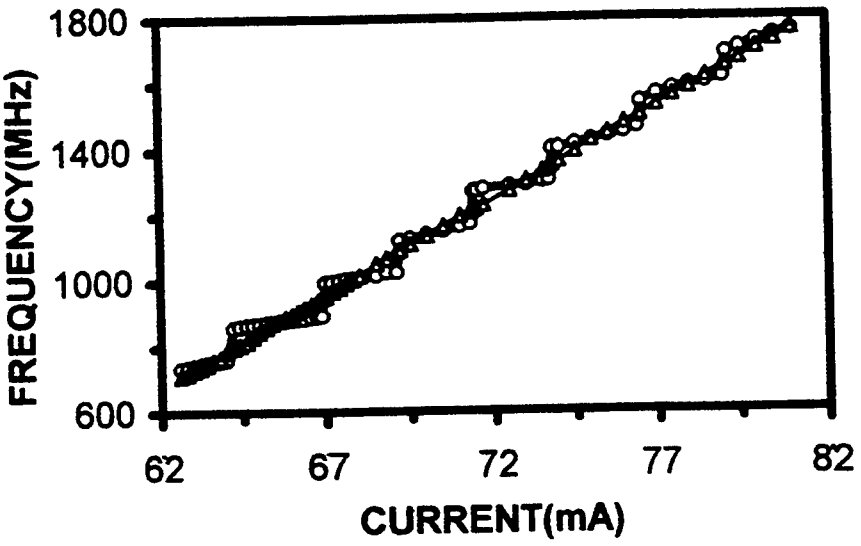


Fig. - 2

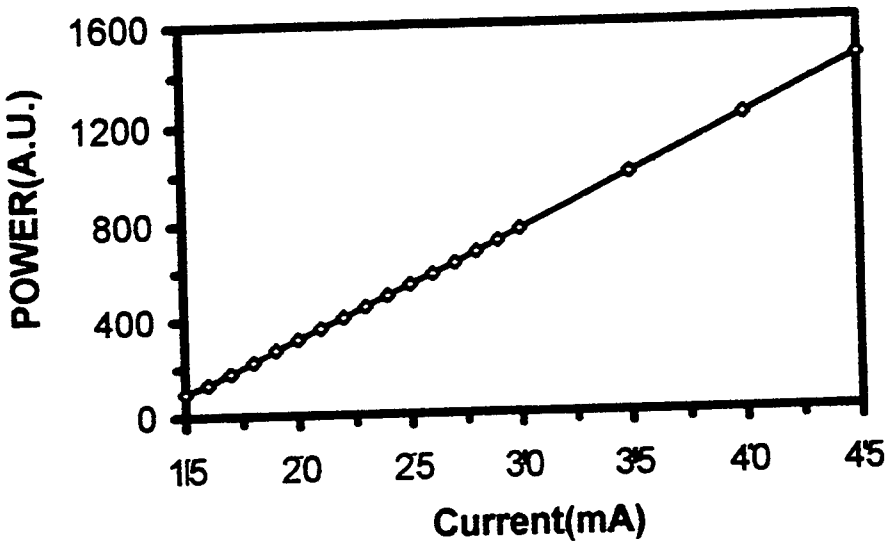


Fig. - 3

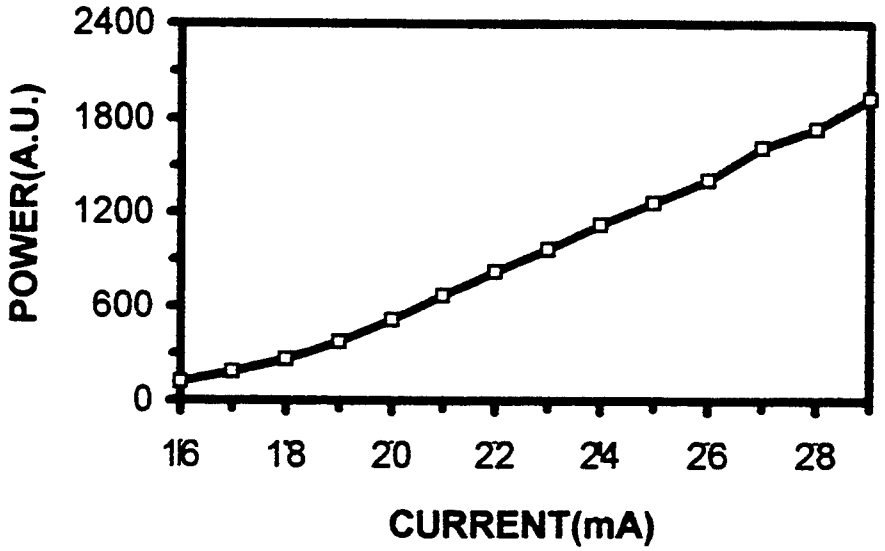


Fig. - 4

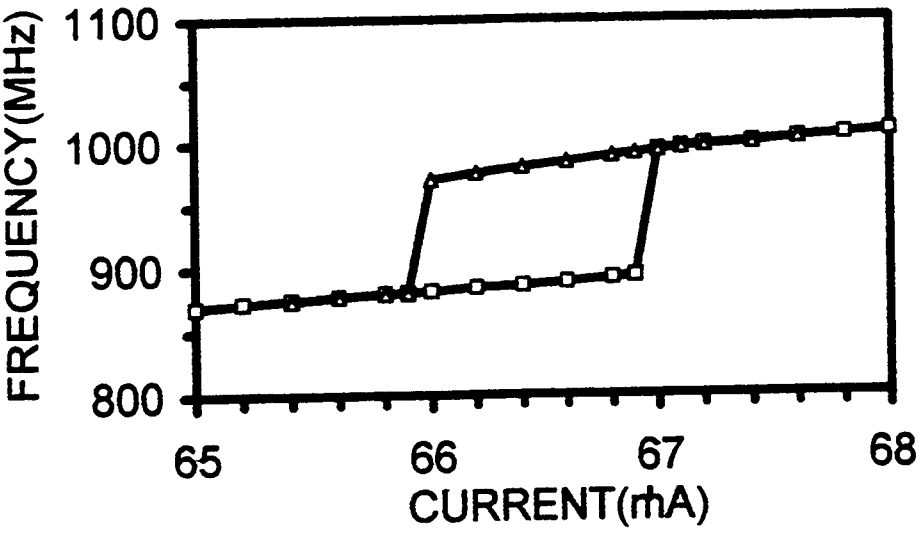

FREQUENCY(MHz)
1100
1000
900
800
65
66
67
68
CURRENT(mA)

Fig. - 5a

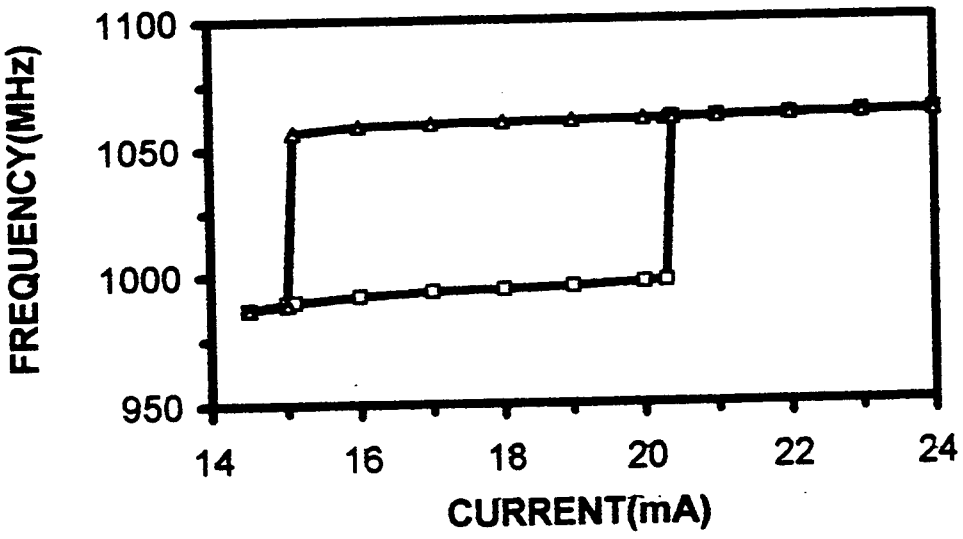

FREQUENCY(MHz)
1100
1050
1000
950
14
16
18
20
22
24
CURRENT(mA)

Fig. - 5b

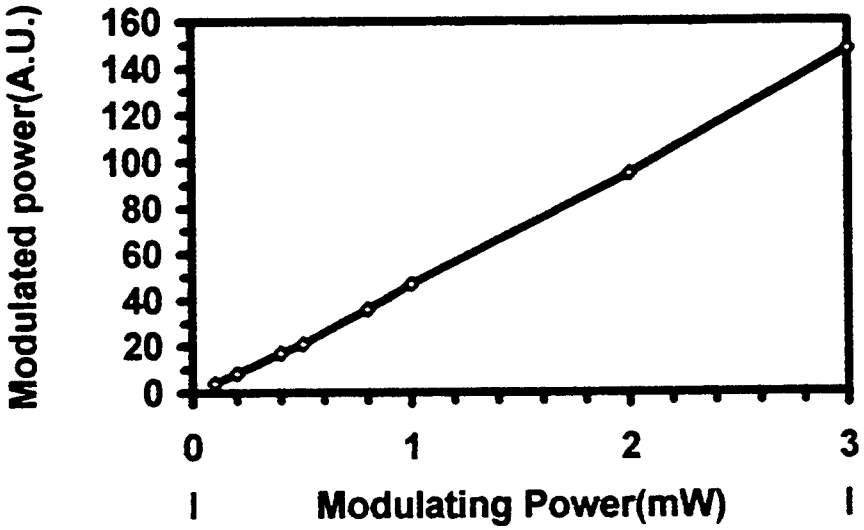

Modulated power(A.U.)
160
140
120
100
80
60
40
20
0
0
1
2
3
Modulating Power(mW)

Fig. - 6

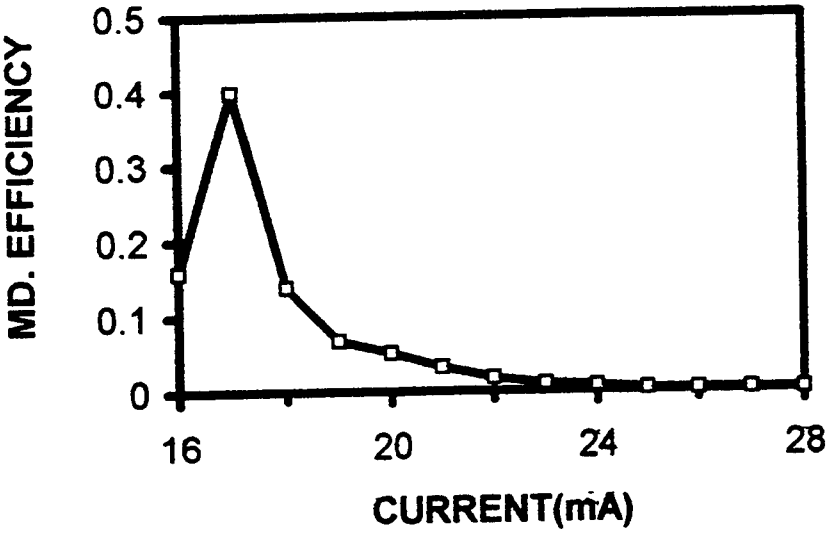


Fig. - 7

EVALUATION OF SEMICONDUCTOR CONFIGURATIONS AS SOURCES FOR OPTICALLY INDUCED MICROWAVE PULSES

Prof. Everett E. Crisman
Department of Physics, Box 1843

Brown University
Providence, Rhode Island 02912

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory
RL/ERAC

September 1996

EVALUATION OF SEMICONDUCTOR CONFIGURATIONS AS SOURCES FOR OPTICALLY INDUCED MICROWAVE PULSES

Everett E. Crisman
Research Professor of Physics
Department of Physics, Box 1843

Abstract

The use of optically induced, DC accelerated, semiconductor carriers as a source of picosecond μ wave pulses is examined. The purpose of this study was to determine 1) whether multiple phase shifted (optical) pulses could be simultaneously generated on a single semiconductor element, and 2) whether two or more, in line, elements could be stimulated with a single optical pulse. Such variations in excitation methods have potential for simultaneously providing the source and phase control necessary for a re-configurable, target recognition, antenna array. The efficacy of both technique are demonstrated in this preliminary study. Also, the gain which could be realized from cooling the semiconductor sources was evaluated for one specimen material. Phase differences for multiple pulses were observed and directly related to the spatial position of the optical pulses on the semiconductor with respect to the μ wave detector. Two cascaded sources, excited with a single pulse, showed enhanced forward μ wave intensity as well as an angular dependence consistent with the double sources and single detector geometry. Finally, cooling from room temperature to 150K resulted in approximately a thirty percent improvement in μ wave strength (from a single source element).

Introduction

Optically excited, semiconductor, photo carriers, accelerated in a dc field, was suggested some years ago by Fattinger and Grischkowsky [1] as a potential source for wide band μ wave pulses in the pico second range. Such sources would have the a time domain width controlled (approximately) by the duration of the optical pulse and the lifetime of the photo carrier species and , if feasible, would allow construction of multiple μ-wave sources arrays without the complexity of wave guides thereby significantly reducing the coupling complexities inherent in the electronics of re-configurable antenna arrays. In addition, the configurations, as described, could act as their own radiating elements (antennae) permitting very compact arrays which could be steered electro-optically [2,3]. Also, since mobility and hence carrier velocity in semiconductors increases with temperature, an improvement in μ wave field strength might be anticipated if the semiconductor source/antenna is reduced in temperature. Cooling, to at least LN_2 temperature, is feasible for both terrestrial and airborne applications.

The initial proof of concept was demonstrated at several laboratories including this one [4]. The concept simply stated, is that photo carriers, induced in a semiconductor by an optical pulse, will accelerate if a dc. electric field is present along and within the semiconductor (say between two surface metallic contacts). Such accelerating carriers (generally electrons) will radiate electromagnetic fields in proportion to the applied dc. field strength up to some maximum velocity controlled by the semiconductor intrinsic parameters. The magnitude of the resulting E-M radiation will be related to the final velocity through the semiconductor mobility and the duration of the pulse will depend on the lifetime of the photo carriers. Experimental results of the past two years have generally confirmed this hypothesis. Also, The dc. field dependence has been reported recently by the Liu, et al. at USAF Rome Lab. Hanscom, MA [5]. In that study, GaAs and InP were examined as a function of applied dc. bias and it was demonstrated that, for both materials, a plateau in the radiated field was reached for the dc. field above some threshold – about 5.5 kV for GaAs and 12 kV for InP. We have expanded on that study to investigate some of the configuration ideals that were suggested by the earlier work cited. The two variations of the basic concept that we evaluate for this study are aimed at providing wide band μ wave pulses separated in both time and space so

that we can determine the potential of this scheme for beam steering and target recognition. In addition we examine the potential for increasing the μ wave field strength by cooling the semiconductor source below room temperature.

Methodology

A.. *Laser induced, picosecond duration, E-M pulses.*

The system used for these studies was essentially the same as the one described by Liu et al. [5] and is reproduced here, with the permission of the authors, as Figure 1. A mode locked, Q-switched YLF laser was used to generate pulses of approximately 80ps duration at 1053nm wave length which can be used directly or frequency doubled to 526.5nm. Suitable dielectric mirrors are used to optimize the energy in the pulses at either the fundamental or the half wavelength laser lines. The semiconductor specimens are biased with a dc pulse of approximately 100μs duration and optical and electronic phase shifting is used to insure the peak voltage of the electronic pulses corresponds to the temporal position of the (picosecond) optical pulses.

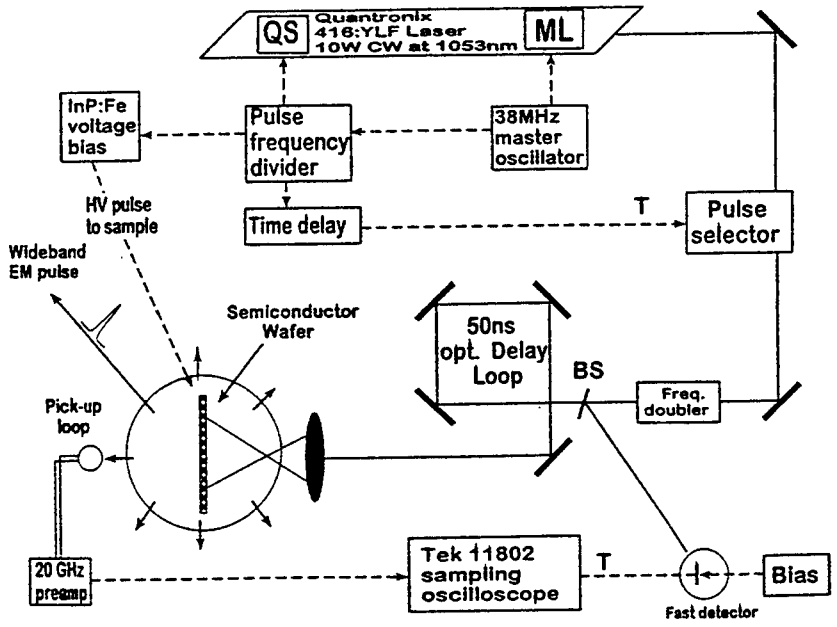


Figure 1: Schematic of the optical and electrical configuration for the experiments described in this report. BS = beam splitter, T = triggered.

A Pockels cell with a 400 Hz repetition rate is used to select the 30 μ Joule, 80 ps duration, laser pulses and a 10:1 beam splitter, combined with an Opto Electronics PD15 ultra fast photo detector, provide the triggering pulses for the Tektronic TEK11802 sampling scope. Fifteen meters of optical path between the triggering detector and the specimen is used to establish a 50 ns time delay compensating for the electronic delay inherent in the sampling scope trigger allowing for the maximum temporal resolution represented by the scope limit of 40GHz. Electric field is applied to the specimen by using an HP 214A generator to supply up to 50 Volt dc. pulses to the low voltage side of a transformer; in this case an automobile ignition coil. Synchronization for the pulsed dc. E-field is taken from the Q-switch of the Quantronix 416 YLF laser and sufficient phases shifting with respect to the fast detector trigger is available in the HP214. Detection of the *radiated* μ wave E-M field is done with a simple 1cm diameter loop used in the near field and various dish and horn antennae for the far field. The relationship between the radiated E-M field and the bias field in the near and far field limits have been discussed elsewhere [6] and will not be repeated here.

The specimens used for this study were all made from auto compensated, CZ grown, GaAs sliced 1/2 to 1mm thick, cut round or rectangular, then chemo-mechanically polished on both sides. Generally, the predominate intrinsic properties sought were short lifetime combined high resistivity: the former to ensure that the carriers do not persist after the optical pulse is removed and the latter to ensure that the bulk material will support the 10-12kV maximum voltage used it the experiments. The contacts, formed by sputtering and annealing Ge:In:Cr:Au, were provided to us by Dr. David Bliss of RL/ERAC [7]. Spacing between the essentially linear contacts was either 1cm or 5cm depending on the particular specimen geometry.

B. The Effect of Temperature on Radiated Field Strength

For this part of the experiment, smaller specimen, 1cm x 3cm, with 1cm space between the electrodes were mounted on the copper cold finger of a glass specimen dewar. Electrical insulating layers were provided between the specimen and the foil heater and between the foil heater and the dewar tail finger. The relatively thick (~ 1mm) alumina used for the latter also provided thermal impedance for the heat with respect to the cold finger. Conversely, the thin (0.25mm) sapphire layer under the specimen afforded

excellent thermal contact to the heater. A resistance thermometer was 'glued' with GE varnish to the sapphire surface adjacent to the specimen which itself was held in place with Dow Corning 340 heat sink compound. Temperature control was automated using Lake Shore Cryogenics, 330 Auto - tuning temperature controller. The specimen dewar vacuum enclosure was modified with two high Voltage glass - to - metal electrode pass throughs and a 22 Meg Ohm surge resistor was mounted inside the vacuum between one of the high Voltage feed throughs and the specimen. Because the high Voltage wires where not routed in through the traditional dewar electrical feed connection, the cold finger shroud was omitted around the specimen. The general layout is shown in Figure 9-2

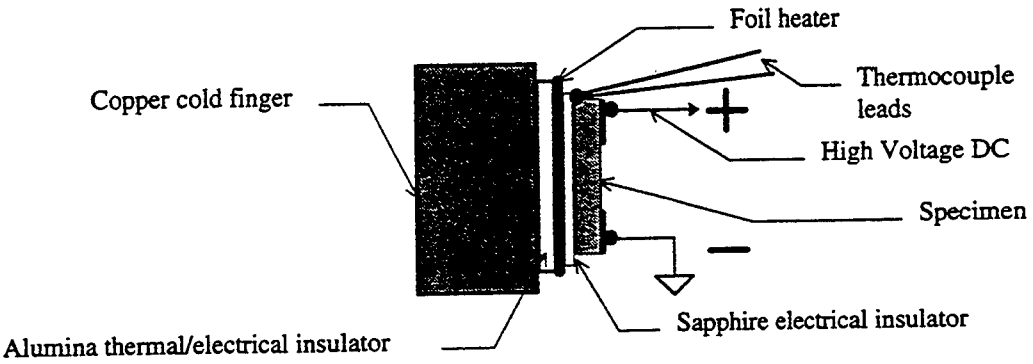


Figure 9-2: Detail of specimen mounting for temperature dependence measurements.

Results

Two configurations were evaluated for this study using the experimental arrangement described above. The first consisted of illuminating a single specimen at two different areas each with 50% of the total beam fluence of the laser pulses. The optical layout for that measurement is shown in Figure 9-3. The laser beams were circular, approximately 6mm in diameter, and spaced 5cm on centers. Some of the salient features of the measurements for that configuration are presented in Figures 9-4 through 9-6. Referring to that set of figures, 9-4 shows the observed double pulse arising from the two spots when their arrivals are displaced in the time domain. Note that the 'inner' pulse in the figure (which came from the beam path *reflected* from the splitter in Figure 9-3) produced a somewhat smaller signal at the detector. This is due to the intensity lost in the extra mirror reflections of the inner beam path. This general relationship between the two photo

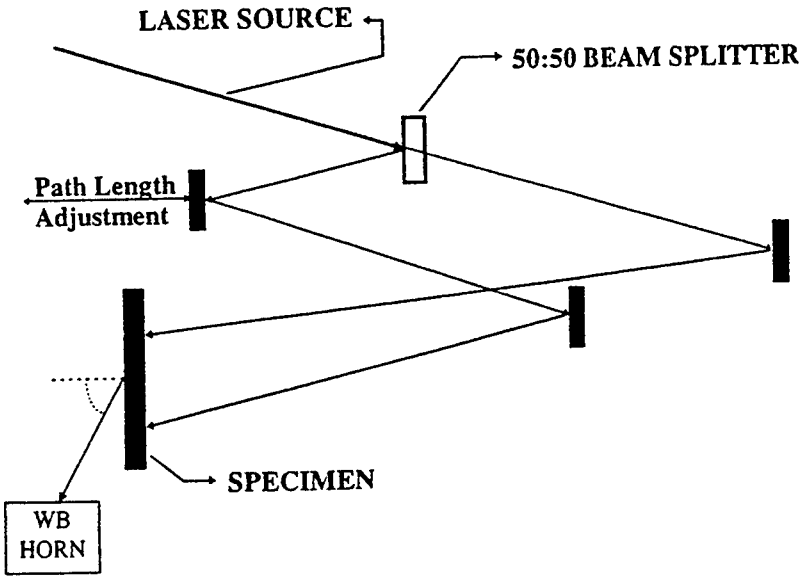
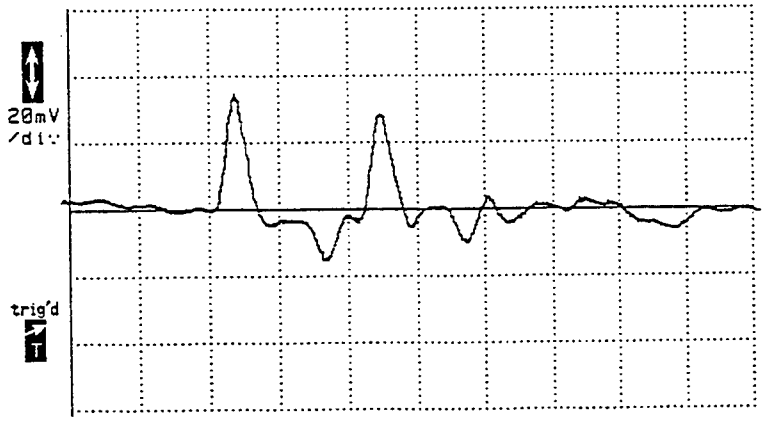


Figure 9-3: Optical configuration for the single specimen/double beam measurements. (WB HORN = Wide band horn μ wave detector.)

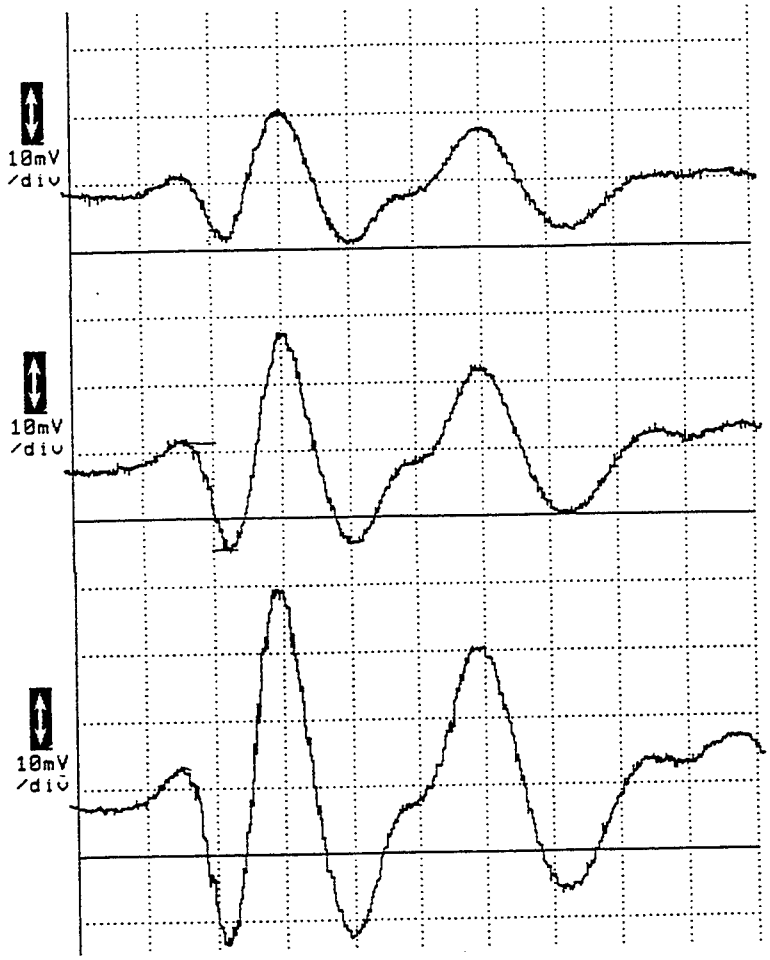
electrically induces pulses was essentially unchanged for spot locations separated either vertically or horizontally on the specimen. The difference in position on the abscissa, about 1ns, is consistent with the 30cm difference in path length after the incident beam was split. As the path length are made equal, the time difference is seen to diminish until the they overlap and the (now single) pulse height just equals the sum of the two individual pulses previously observed (not shown in these figures).



500 ps/div

Figure 9-4: Double radiation pulses emitted by a <u>single</u> GaAs specimen illuminated by two differently delayed and spatially separated laser pulses. Near field measurement.

The trace of Figure 9-4 was taken in the near field. However, the far field is where such systems will be applied and so some additional measurements were done for that condition as well. For the far field condition on axis (i.e. centered on the normal to the specimen surface) a single signal was observed for the



"inside" beam only

"outside" beam only

both beams together

Figure 9-5: Far field signals from double spot illumination of single GaAs specimen. Normal to specimen surface.

double illumination condition with the amplitude being the sum of the two individual amplitudes the two spot sources. The amplitude and peak position on the abscissa were essentially independent of the position of the two spots on the specimen surface. In figure 9-5 the monitored signals from the individual spots and from combined signal from both are shown.

Space does not permit a detailed analysis of the full pattern shown. Most of the oscillations to the right of the first transient are due to reflections of the pulse which travel in directions other than directly to the far

field detector and others are due to ringing of the horn antenna. The first oscillation on the left of the traces represents the signal from the specimen at the antenna which is actually proportional to the *derivative* of the E-field of the μ wave emission. Note that the "inside" spot source intensity is again noticeable small than the others due to the differences in optical fluences of the two spots on the source specimen.

This measurement was repeated for the far field condition but with the detector horn antenna center line rotated in steps from its original position ending at 90 degrees off the normal to the specimen surface i.e. at a right angle to the optical path. A pair of the traces from the sequence are shown in Figure 9-6. for the condition of 90° rotation of the far field detector.

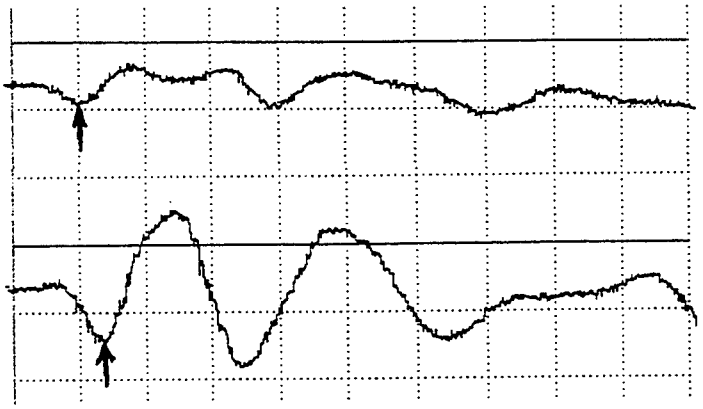


Figure 9-6: Far field signals from double spot illumination of <u>single</u> GaAs specimen. At 90 degrees to specimen specimen normal.

While alignment problems have reduced the total signal strength for this measurement to the point where reflections and detector ringing dominate, the most important feature can be seen by comparing the temporal positions marked with arrows on the figures at the derivative minimums. The time shift about 50 ps is consistent with the 5cm distance difference from the detector to the two illumination spots! Thus it is possible to spatially resolve multiple spots illuminations <u>on the same specimen</u> as a function of the angle of measurement off the system axis.

The optical arrangement for the single beam / dual source measurements was similar except that that the two beam were reflected to the backs of two separate specimens each with its own dc. pulse source. The

configuration for the specimen positioning is shown in Figure 9-7.

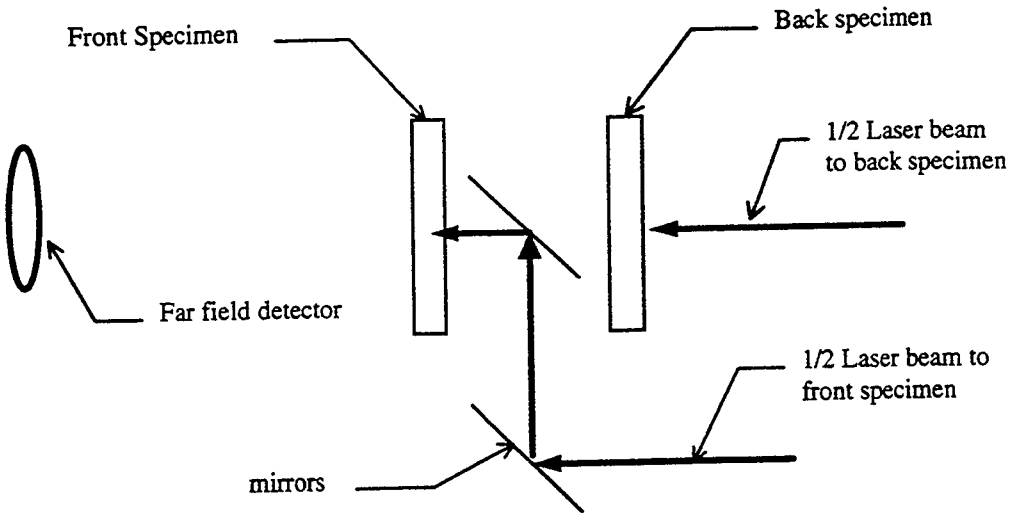


Figure 9-7: Optical arrangement for two specimens excited by a single optical source

Once again the space limitations of this report do not permit a complete analysis of the measurement to be presented. However the most important detail of the investigation is shown in Figure 9-8. As in the two-beam/single-source set up, it was possible to demonstrate for this scheme as well that the forward μ wave pulse is a composite of the two pulses from the individual specimens. Note that there is a slight miss-alignment in the peaks with respect to time indicating that the path lengths of the split sources was slightly longer for the front beam. Also the composite signal of the double source appears narrower indicating, perhaps, that signal superposition is resulting in a wider band source! While not explicitly discussed in this report, it was also demonstrated as part of these experiments that the amplitude of the μ wave pulses from each source could be **varied independently** by varying the dc. pulse voltage applied to the specimen contacts. Also we determined that the insertion of a piece of 1mm, pyrex, flat glass, with 500A of aluminum, as the final mirror in the front specimen optical path introduced neglectable attenuation on the far field μ wave signal strength from the back specimen source. As of this report, data had not yet been accumulated, in this configuration, for far field μ wave signal strength as a function of angular position of the detector.

The final measurements of this report where done to determine the temperature dependence of the μ wave

field strength as a function of specimen temperature. Because the dewar window was only 1" in diameter near field measurements only were done for this investigation. As stated above, the presence of the high voltage leads and the current limiting resistor within the small vacuum chamber did not permit the thermal

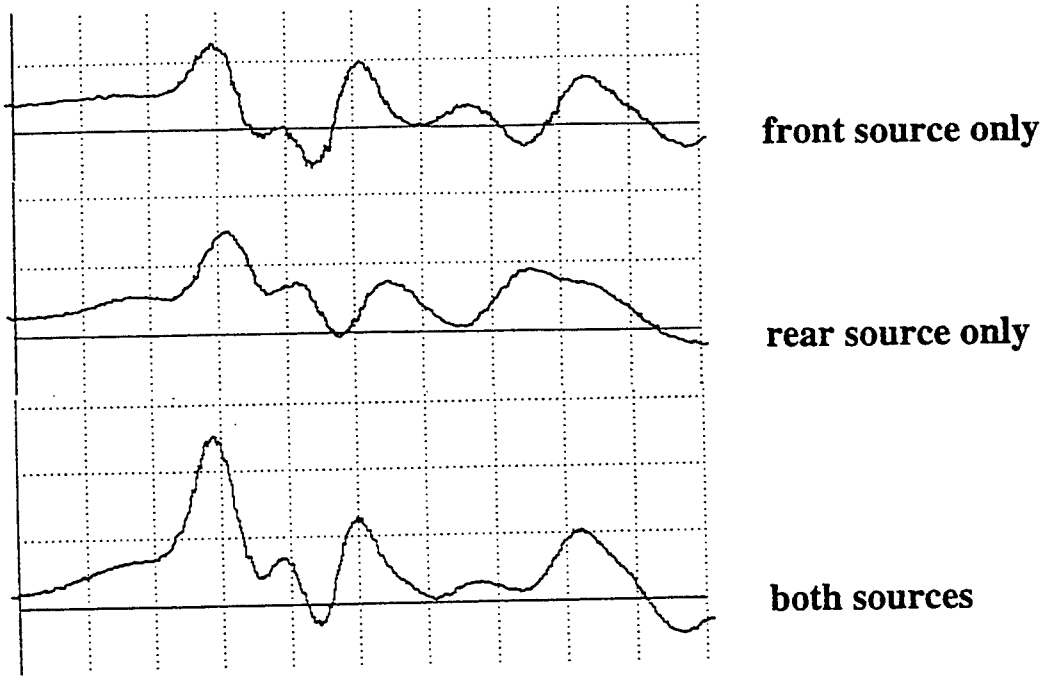


Figure 9-8: Far field measurements of the single beam and cascaded double specimen configuration.

shrouding of the specimen from the thermal radiation of the surroundings. Therefore, it was impossible to attain even the LN_2 temperature limit of the dewar itself. Nevertheless, data taken from room temperature down to 150K showed that the field strength was increasing inversely as the temperature decreased. Further when this was compared to the measured value of electron velocity versus temperature [8] the slope of the two was identical. The increase was observed to be 30% for a chance in temperature from 300K to 150K and appears to be entirely due to the increase in the electron velocity. A graph of these results is shown on Figure 9-10.

Discussion

We have demonstrated that two semiconductor sources can be simultaneous excited by optical pulses to produce algebraically additive μ wave fields at a distance from the generation point. Thus the sources may

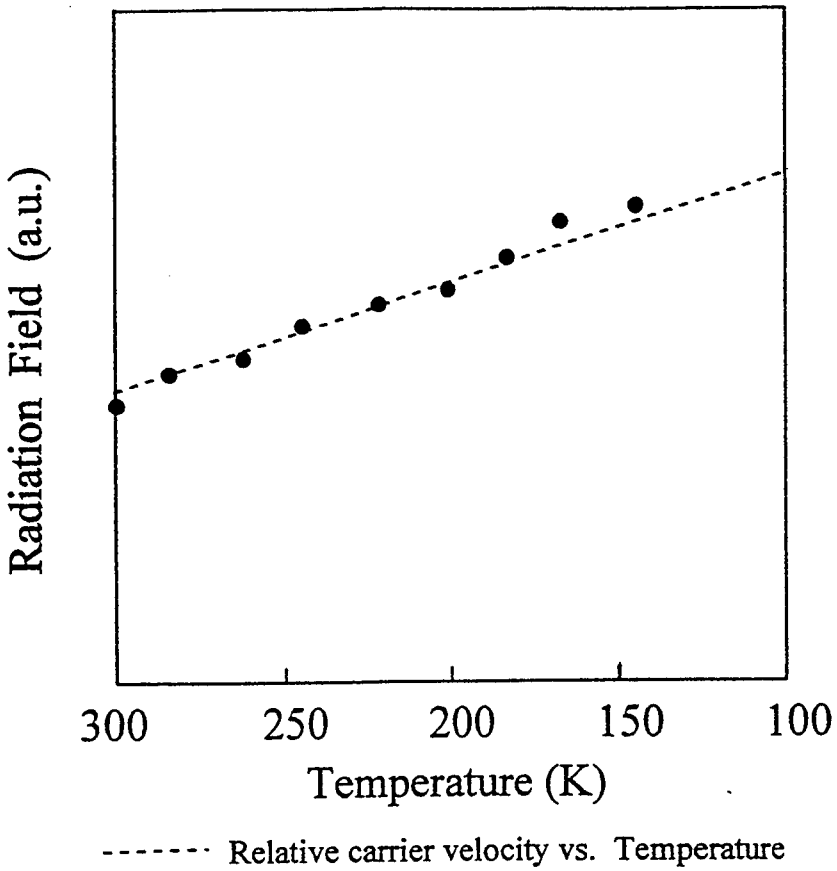


Figure 9-10: Strength of the far field μ wave signal versus temperature for a single GaAs specimen and single laser beam as a function of specimen temperature. The dashed line is plotted from data in [8].

also be interpreted as the antenna elements in an array. Variations of this approach should allow optical reconfigurablity of an array without complex electronic switching or impedance matching. In addition, some gain can be realized by reducing the temperature of the semiconductor sources. Although not discussed here, experiments into the field strength an a function of laser wave length and semiconductor properties were also begun, during the course of this project, which have great potential for further increasing the μ wave far field strength. Two submission based on this work were prepared and one has been presented. The abstracts are appended to this report.

Acknowledgments:

The author thanks Dr. David Liu for the great deal of time he spent in discussion of the physics of this

Acknowledgments:

The author thanks Dr. David Liu for the great deal of time he spent in discussion of the physics of this project, instruction on the optical system and for his aid in setting up the system and acquiring the data. "Thank you" to Dr. John Derov of RL/ERAC for many insightful discussions and comments and to J. A. Riordan for preparing some of the contacts for the specimen use here. Finally a particular thanks to Dr. Steven Mittleman of RL/ERAC for inviting me to join his group for this summer. No short research project such as this can be accomplished without input from a great many people at the organization visited. So I also thank those whom space does not permit me to name individually!

References

1. Ch. Fattinger and D. Grischkowsky, **Tetrahertz Beams,** *Appl. Phys. Lett.*, Vol. 54, pp 490 - 492, (1989).
2. B .B. Hu a. t. Darrow, X. -C. Zhang and D. H. Auston, **Optically Steerable Photoconducting Antennas**, Appl. Phys. Lett., Vol 56, No. 10, pp. 886 - 888, (1991).
3. X. -C. Zhang and D. H. Auston, **Generation of Steerable Submillimeter Waves from Semiconductor Surfaces by Spatial Light Modulators**, Appl. Phys. Lett., Vol. 59, pp. 768- -770 (1990)
4. D. W. Liu, J. B. Thaxter and D. F. Bliss, **Gigahertz Planar Photoconducting Antenna Activated by Picosecond Optical Pulses,** *Optics Letters*, Vol. 15, No. 14, pp. 1544 - 1546 (July, 1995).
5. D. W. Liu, P. H. Carr and J. B. Thaxter, **Nonlinear Photoconductivity Characteristics of Antenna Activated by 80 - Picosecond Optical Pulaes,** IEEE Photonics Technology Letters, Vol. 8, No. 6 (June, 1996).
6. See for example: E. Marengo **Theoretical Aspects of a GaAs Photoconducting Antenna Excited by 80ps Laser Pulses**, this publication.
7. J. A. Riordan, **Experimental Study of Rogowski Profile InP and GaAs Wafers**, this publication.
8. See for example: S. M. Sze, *The Physics of Semiconductor Devices,* 2nd ed., Wiley New York, NY (1981)

MULTIDIMENSIONAL OPTICALLY EXCITED PHASED ARRAY PHOTOCONDUCTING ANTENNAS

D. W. Liu*, E. E. Crisman#, J. S. Derov, P. H. Carr, and S. D. Mittleman

Rome Laboratory
Electromagnetics and Reliability Directorate (RL/ERAC)
Hanscom AFB, MA 01731-3010
617-377-4038 voice, -1074FAX
mittleman@maxwell.rl.plh.af.mil

*Current Address: Sanders, A Lockheed Martin Co. 95 Canal Street, Nashua, NH 03061
#Permanent Address: Department of Physics, Brown University, P.O. Box 1843, Providence, RI 02912

Photoconducting antenna elements activated by picosecond laser pulses are implemented as 1~20 GHz reconfigurable electromagnetic radiation sources. The microwave radiation generated by the dc driven photocurrent is detected at the near field with an inductive loop and at the far field by an impulse antenna. The laser pulse at the infrared wavelength region (less than the bandgap energy of the semiconductor antenna element) is employed as the optical resource, which allow the laser beam to be partially transmitted through and partially absorbed in the sample. Such a arrangement makes a 3-D antenna phased array configuration possible. The elements in a series configuration also can be excited by optical sources fed from the side in a synchronous manner. In order to depress the mutual disturbance or coupling between the elements, the bias fields of the elements are set at the plateau region [D. W. Liu, P. H. Carr, and J. B. Thaxter, "Nonlinear Photoconductivity Characteristics of Antenna Activated by 80-picosecond Optical Pulses", IEEE Photonics Technology Letters, Vol. 8, 815-817 (1996)]. As a result, the generation of the microwave pulse from the subsequent element will not be affected by the presence of the microwave pulse from the first element.

We also observed that the microwave pulse profile is dependent on the wavelength of the excited optical pulse. Consequently, it is possible to amplitude modulate the microwave pulse profile by varying the wavelength of the laser pulse or by physically rotating the antenna element. Some numerical analysis associated with this unique capability will be presented.

Abstract submitted for publication to "The DRAPA Symposium on Photonics for Antenna Applications", January, 1997, Monrterey California.

IMPULSE PHOTOCONDUCTING ANTENNAS ACTIVATED BY PICOSECOND LASER PULSES

D. W. Liu, P. H. Carr, E. E. Crisman*, and D. F. Bliss

Rome Laboratory
Electromagnetics and Reliability Directorate (RL/ERAC)
Hanscom AFB, MA 01731-3010
617-377-3581 voice, 617-377-1074 FAX
dwliu@maxwell.rl.plh.af.mil

* Permanent address: Department of Physics, Brown University, P. O. Box 1843, Providence, RI 02912

ABSTRACT

Reconfigurable microwave impulse antennas can be implemented optically. In this work, 50 - 100 picosecond laser pulses from a frequency-doubled, mode-locked, Q-switched YLF laser generate photoelectrons in dc-biased high resistivity semiconductor wafers. We have investigated InP:Fe, GaAs, and Low -Temperature grown GaAs (LTG - GaAs) for this application. The microwave radiation due to the dc-driven photocurrent is detected at the near field with an induction loop and at the far field by an impulse antenna. These signals are observed in real time with a Tektronix 11802 sampling oscilloscope. We have studied nonlinearities in the microwave radiation for optical fluence as high as 300μJ/cm^2. Nonlinearities at dc-bias fields as low as 12KV/cm were also observed. The relevant parameters involved for generating microwave pulses, such as photoconductivity, bias field strength, optical fluence, antenna element area, and experimental observations will be analyzed and discussed.

Since the output microwave pulse profile emulates the input optical signal, pulse-shaping can be implemented by modulating the optical beams. Multi-element phased array antennas can be established with multiple laser beams or with fiber optic feeds for individual elements. Because the 1.06μm excitation beam will be partially absorbed and partially transmitted through the semiconductor antenna element, 2-D and 3-D array are possible. We will also describe the results of a 3-D serial configuration as a proof of concept for this antenna.

Abstract of paper presented at the "1996 Phased Array Antenna Symposium", September, 1996, Allerton Park, Il.

TECHNIQUES FOR DETERMINING THE PRECISION OF RELIABILITY PREDICTIONS AND ASSESSMENTS

Digendra K. Das
Associate Professor
Department of Mechanical Engineering Technology

SUNY Institute of Technology at Utica/Rome
Marcy Campus
P.O. Box 3050
Utica, NY 13504-3050

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

August 1996

TECHNIQUES FOR DETERMINING THE PRECISION OF RELIABILITY PREDICTIONS AND ASSESSMENTS

Digendra K. Das
Associate Professor
Department of Mechanical Engineering Technology
SUNY Institute of Technology at Utica/Rome

Abstract

A preliminary investigation of the various techniques available for determining the accuracy of reliability predictions was undertaken. The research project was designed as a complementary effort in support of the "New System Reliability Assessment Methods" program sponsored by Rome Laboratory N.Y. Classical statistical techniques used in probability theories were explored. The applicability of alternative approaches using possibility theory was investigated. Also the development of practical user-friendly reliability assessment techniques was studied.

TECHNIQUES FOR DETERMINING THE PRECISION OF RELIABILITY PREDICTIONS AND ASSESSMENTS

DIGENDRA K. DAS

INTRODUCTION

It is well known that there exists a significant difference between predicted reliability and actual field reliability of systems. This is often termed as "Reliability Delta" and has been a long standing problem in reliability engineering. A review of the published literature indicates that a considerable amount of research effort has been directed to identifying the factors contributing to the reliability delta.

Usually system reliability is derived from the knowledge of component reliability and the concepts of statistical life-time distributions. However, in practice this knowledge does not necessarily provide insight into how a single component will behave in a system. As a consequence systems all too often achieve reliabilities markedly different (usually lower) than those predicted.

In response to this long standing problem, Rome Lab has initiated the "New System Reliability Assessment Methods" program. The objective of this effort is to develop a system reliability methodology that accounts for all predominant factors that affect field reliability of systems. The performing organizations in this program are IIT Research Institute/Reliability Analysis Center (IITRI/RAC) and Performance Technology.

The current Summer Faculty Research Program (SFRP 1996) has been designed to be a complementary effort to the on going "New System Reliability Assessment Methods" program.

SUMMER FACULTY RESEARCH PROGRAM

OBJECTIVE

The objective of this Summer Faculty Research Program was to initiate the development of practical methods for determining the accuracy of reliability predictions and assessments. The methods will be used to estimate the precision of reliability assessments that take place during the life cycle phases of system development.

SCOPE

The scope of the effort consists of the following:

I) Familiarization with the ongoing "New System Reliability Assessment Methods" program.

ii) Review of the literature on related research.

iii) Study of the application of classical statistical methods to estimate variance or confidence intervals on reliability predictions and assessments.

iv) Study of the applicability of information/system theory and other available theories to estimate the accuracy of reliability predictions and assessments.

v) Study of practical techniques and methodologies available for reliability predictions/assessments and determination of the accuracy of these tools.

BACKGROUND RESEARCH (TASK 1)

This section gives a brief outline of the "New System Reliability Assessment Methods" program and its current status. The details of the program are available in References 1 & 2.

As mentioned earlier, this program is sponsored by Rome Laboratory and performed by the IIT Research Institute/Reliability Analysis Center (IITRI/RAC) and Performance Technology. The objective of the program is to develop a systems reliability methodology that accounts for the predominant factors that effect field reliability and all types of failures. The types of failure identified for inclusion in the development of the reliability model are:

- Part Defects
- Design Defects
- Excess Stress
- Fatigue
- Parametric Drift

- Manufacturing Defects
- Assembly Defects
- Part variability
- Manufacturing Variability
- Human Factors
- Interactions
- Embedded Software

In order for the model to be successful, it should assess reliability as a function of:

- Adequacy of Requirements Definition
- Part/Material Adequacy
- Design Robustness
- Manufacture/Assembly Process Adequacy
- In Process Defect Data
- Test Data

This methodology would assess the reliability of a system based on the best possible data/information available at the time. Other advantages are that:

- It accounts for all factors than can effect reliability in field use.
- It uses the best available data and does not require specific or complete data sets.
- The resulting assessments will be more realistic since they are based on empirical data on the same or a similar system.
- It can assess a manufacturer's ability to design and build a reliable product.
- It provides factors for determining operational reliability.
- It grades efforts to improve reliability.

The program consists of six tasks to be completed in three phases. The tasks are:

Task 1: Identify and analyze initial reliability assessment methodologies.

Task 2: Gather information on the purpose and methods of current practices.

Task 3: Investigate potential methodologies for assessing system design and manufacturing processes.

Task 4: Investigate potential methodologies for improving the prediction with empirical test data.

Task 5: Methodology Development

Task 6: Methodology Validation.

Phase 1 of the program, which includes Tasks 1 through 4, is already complete and reported in reference 1. Phases 2 and 3 are now in progress and expected to be completed by late 1997. These phases will comprise of the remainder of Tasks 3 and 4 which were initiated in Phase 1 and the last two tasks.

During Phase 1 of the program a literature search was conducted to identify data and information pertinent to the project.

Attention was given particularly to identifying the sources of information on system failure modes and modeling methodologies. 135 documents and papers were identified and reviewed; 39 on Assessment Methodologies, 21 on Failure Mode/Mechanism Data, and 75 on various pertinent methodologies.

The study also attempted to ascertain the purpose for performing reliability predictions and the manner in which they were performed. The sources of information used in this part of the study were:

I) A survey of the reliability professionals

ii) A previous RAC study entitled "Benchmarking Commercial Reliability Practices".

iii) Solicitation of technical opinions/practices from various organizations.

iv) Documentation of technical inquiries received at RAC.

It was found that the predominant purposes for performing reliability predictions were:

1. Determining feasibility in achieving reliability goal or requirement
2. Aiding in achieving a reliable design (i.e., derating component selection, environmental precautions, input to FMEAs/Fault Trees)
3. Predicting warranty costs and maintenance support requirements

It was also determined that MIL-HDBK-217 (Ref 13) was the most universally applied failure rate reference manual and that customized versions of MIL-HDBK-217 could result in accurate system level reliability assessments. Several organizations reported that their assessments were within 5-15% of the observed field failure rate.

As indicated earlier, phases 2 and 3 of the "New System Reliability Assessment Methods" program are now in progress. Based on the results of Phase 1 of this project a new assessment methodology entitled "Consolidated Reliability Assessment Methodology (CRAM)" has been proposed. The CRAM model is a

comprehensive approach to reliability prediction and estimation. The following advantages of the model have been identified:

1. It will focus on process metrics as an adjunct to predicting field reliability. This is in concert with today's process emphasis strategies such as ISO 9000, Malcom Baldrige, and Software Engineering Institute evaluations. Process measures are timely to manage and preclude having to wait for the final product to be seen before its quality can be ascertained. It will also enable mid-course corrections. It will be updated throughout the development cycle and, thus, will dynamically represent the state of the system.

2. The model will influence the design in real time. It can be integrated into the design, development and manufacture of the product. It will provide guidance to achieving better reliability. This is an important advantage that most other reliability models cannot provide, at least to the extent envisioned for the projected model.

3. It will be a model that distributes responsibility throughout the organization. All people in the organization will see the impacts that their defect data will have on the projected reliability. They will also have a menu of tools that they can consider applying to improve the quality of their product.

4. It will provide guidance to improve reliability, so that designers and others will know how to best improve the predicted reliability of their designs.

5. The model will link together tools, process and metrics to generate failure rate predictions. This model will be a reliability estimator that will also serve as a catalyst to bring a common development focus on product and process reliability.

The planned tasks for the development of the new CRAM model are:

1. Identification and quantification of process factors that determine the operational reliability.

2. Grading the impact that process tools will have on the assessed reliability.

3. Further development of in-process development measures to predict the latent faults at time of shipment, and then to transform those into operational reliability measures.

4. Further development of a Consolidated Reliability Assessment Method model.

5. Calibrate and validate the methodology.

6. Harmonize the new reliability assessment method terminology, format, and operation with co-existing Rome Laboratory reliability and quality models.

APPLICABILITY OF STATISTICAL METHODS (TASK 2)

The current Summer Research program was designed to be a complementary effort to the "New System Reliability Assessment Methods" program. It was established in consultation with IITRI/RAC and Performance Technology that an initial reliability prediction method (probably MIL-HDBK-217) would play a role in The New System Reliability Assessment method currently under development. The initial prediction will then be refined as data becomes available during system design and development. The intent is to combine initial predictions with the best available data for the purpose of improving the accuracy of the reliability figure of merit. It is therefore essential that a study of the techniques for determining the precision of Reliability prediction and assessment models be undertaken.

A literature search was conducted using the Rome Lab and IITRI/RAC library facilities to identify techniques related to variance/confidence intervals and reliability estimation/prediction. The data base research was conducted to cover a period of about 35 years (1960-1995) and a total of 151 documents and papers were identified and reviewed. Some of the documents relevant to the current project are listed in the reference (Ref 3-20). Several techniques for determining the precision of reliability predictions and assessments are available in these published works. Judnick (Ref 3) suggested that in order to obtain an exact estimate of system failure rate one should use the Mellin transform. A formal definition of the Mellin transform is available in reference 40. Unfortunately, the mathematics involved to obtain these results are not accessible to all. The likelihood of error in the tedious calculations is considerable even in the case of the two-component system. Hence to obtain approximate estimates of reliability, simulation methods are probably more suitable. However, to get a quick first-cut rough estimate, Rosenblatt's method (Ref 18) can be used.

The exact Mellin transform method for Bayesian Confidence intervals was published by Spronger and Thampson (Ref 19) in 1966. Approximately about the same time Levy and Moore (Ref 20) showed the broad applicability and advantages of a Monte Carlo simulation approach to the problem of reliability predictions and

assessments. There are now many published articles available in the modern literature, addressing these problems.

One of the objectives of the Summer Research Project was to identify and obtain reliability data which can be used to study the application of classical statistical methods to estimate variance or confidence intervals of reliability predictions. The information available in Ref 12 was chosen for this purpose. The data was obtained and procedures were initiated to estimate the precision of MIL-HDBK-217E reliability prediction models for capacitors. The objective of the study documented in Ref 12 was to update the MIL-HDBK-217 failure rate prediction models for Capacitors, Resistors, Inductive Devices, Switches, Relays, Connectors, Interconnection Assemblies and Rotating Devices. These models were developed or modified primarily from the statistical analysis of field failure data collected during the study. All reliability models relied on field data except for interconnection assemblies which used laboratory test data. The methodology used for the models essentially converts a time to failure statistic such as Mean-Time-to-Failure (MTTF) or characteristic life to an average failure rate over the design life cycle or preventive maintenance interval. Since a closed form solution for the calculation of this average failure rate is not possible, it was accomplished by means of Monte-Carlo simulations.

The reliability models developed in the study were based on Linear Multiple Regression Analysis data for components such as Capacitor, Resistor etc. A typical set of data from this analysis has been shown in Table 1.

Based on the preliminary work completed in this task of the Summer Research program a follow on research project has been planned in consultation with IITRI and Rome Lab. The project will involve the development of a confidence level simulation model based on data available from Reference 12 (Parts count method) and extending the model to circuit level systems. It is anticipated that the proposed simulation model will then open up the possibility of integrating the model with the upcoming CRAM model. (Ref 1 & 2).

TABLE 1
Capacitor's Data (Ref 12)

Normalized to: Fixed Paper Capacitor
Ceramic Package
G_B Environment
Operating Environment

VARIABLES IN THE EQUATION

Variable	B	SE B	95% Confidence	Interval B	Beta
D9 (ta elec)	-1.69487	.35541	-2.39621	-.99354	-.19337
D7 (plastic)	1.18050	.50714	.17975	2.18125	.10396
D6 (mica)	.89810	.50994	-.10817	1.90438	.10749
D4 (electrolytic)	-.62003	.54582	-1.69709	.45703	-.08549
D3 (ceramic)	-.58884	.56212	-1.69807	.52040	-.08119
E6 (A_{UF})	6.27433	.54189	5.20502	7.34364	.67734
E5 (A_{UA})	5.31034	.56029	4.20472	6.41596	.53730
E3 (A_{IC})	7.34207	1.04160	5.28668	9.39747	.46586
P4 (metal	-1.46382	.91764	-3.27461	.34698	-.07076
package)	2.08365	.56702	.96475	3.20256	.15677
F1 (variable)	2.75466	.48584	1.79595	3.71337	.41869
E4 (A_U)	4.24595	.60044	3.06109	5.43080	.47600
E8 (G_F)	-4.71904	.77249	-6.24341	-3.19467	-.35504
TT1 (nonop)	8.13377	1.45071	5.27108	10.99647	.27947
E3 (A_{IF})	-1.57215	.92108	-3.38973	.24544	-.12372
P5 (plastic	-2.43752	1.62391	-5.64199	.76694	-.05937
package)	-1.21026	1.08532	-3.35192	.93140	-.04158
D1 (air)	-18.87137	.22402	-19.31344	-18.42931	
E7 (G)					
(Constant)					

Multiple R	.89146		
R Square	.79470	R Square Change	.00143
Adjusted R Square	.77520	F Change	1.24349
Standard Error	1.38686	Significant F Change	.2663

F = 40.75764 Significant F = 0.0

APPLICABILITY OF INFORMATION AND OTHER AVAILABLE THEORIES (TASK 3)

One of the objectives of the Summer Research program was to study the feasibility and practicality of applying the General System and other available theories to estimate the accuracy of reliability predictions and assessments. The reason for identifying this task (alternative approaches) as one of the primary objectives of the summer program is as follows:

For a long time, the probabilistic approach to system reliability has been found to be adequate and highly useful in assessing the performance of various systems. However, the major shortcoming of this approach lies in its failure in offering an effective tool to handle the problem of uncertainty. The uncertainty of results puts the entire process of system reliability assessment to question. Specifying the mean and variance or confidence levels is not enough to tackle the problem of uncertainty. The Bayesian models in probability theory, although widely used as a numerical approach for representation and inference with uncertainty, mask the problem of uncertainty caused by the ignorance remaining hidden in the priors.

The limitations of the reliability assessments based on the probability theory becomes apparent, if we note some of the known worst accidents in human history, e.g., the accidents of the nuclear power plants at Three Mile Island (1979) and Chernobyl (1986), the explosion accident of the Challenger (1986), a crash accident of a Japan Air Lines jumbo jet (1985), and an accident at a chemical plant at Bhopal (1984). These accidents were deemed highly unlikely on the basis of the probability theory since their probability of occurrence was estimated to be very very small. It should therefore be noted that a small probability for an event does not always mean low possibility of the event.

An extensive literature search was conducted to identify alternative approaches. Some of the most relevant documents and papers are listed in the Section on Reference (Ref 21-32). The theories explored are:

1) General System Theory (Ref 21-23)
2) Chaos Theory (Ref 24)
3) Fuzzy Sets Theory (FST) (Ref 25-27)
4) Dempster-Shafer Evidence Theory (DS/ET) (Ref 25,28,29).

It became apparent at the onset of this research project that the most promising theories relevant to reliability assessment would be The Fuzzy Sets Theory (FST) and Dampster-Shafer Evidence Theory (DS/ET). Due to the time constraints of the summer project (Maximum 12 weeks), General System and Chaos Theories were not pursued in detail. An outline of the FST and DS/ET and their applicability to reliability assessments is presented below:

FUZZY SETS THEORY (FST)

This theory was originally presented by Zadeh (Ref 26,27) and provides the basis of a possibilities approach to system reliability evaluation based on the premise that a small probability does not always mean a low possibility of an event, whereas a low possibility would necessarily imply a low probability (Ref 25).

FST may be thought of as a generalized form of the Conventional Boolean Set Theory. The difference is that an object can have a fuzzy set membership anywhere in the continuous range from 0 to 1, but for Boolean sets membership is restricted to values exactly equal to 0 or 1. The Fuzzy set operation procedures (Union, intersection...) are defined such that they reduce to the corresponding Boolean expressions when they are applied to sets with memberships restricted to 0 and 1. Similarly, fuzzy logic operations (AND, OR,....) reduce to their Boolean equivalents at the boundaries of their domain (Ref 32).

DEMPSTER-SHAFER EVIDENCE THEORY (DS/ET)

The DS/ET originated from the work of Dempster (Ref 28) and was developed by Shafer (ref 29). It is a new tool for representing situations in which various kinds of ignorance exist in our knowledge or information about a system. It is a reasoning approach for testing multiple hypothesis on the basis of evidence. The only assumption made about the evidence is that the sum of the values supporting all possible conclusions plus the unknown equals 1. Evidence from multiple sources is combined by a geometric procedure that generates the mass of evidence supporting each possible conclusion.

APPLICATION OF FST AND DS/ET IN RELIABILITY ASSESSMENT

A review of the modern literature in this area indicates that the FST and DS/ET have been successfully applied in the area of system reliability evaluation (Ref 25). In a recent project (Ref 32) these theories along with the well established probability

theory were applied successfully to establish the confidence level in reliability of a missile system. The Expanded Confidence Assessment Process (EXCAP) model developed in the project used the principles from FST to transform raw test results into "evidence" parameters and to assist in utilizing dissimilar information. Bayesian inference nets, an application of Bayesian statistics, were used to provide a structure or network for propagating evidence from lower levels of the system into effects on the total missile system. Finally DS evidential reasoning permitted the incorporation of evidence which is partially relevant, providing an accounting of the amount of "Unknown" associated with its use and a generation of confidence bounds which reflect supporting and conflicting evidence.

DEVELOPMENT OF PRACTICAL TECHNIQUES (TASK 4)

The new techniques and developments identified in the previous sections show that the upcoming new methologies will be highly sophisticated in nature when fully developed. It is quite obvious that to use these new methologies, the users must also have a higher level of intellectual capability. This is certainly a very significant disadvantage in practice for non-expert or non-specialist users. The aim of this task was to outline the framework for a practical and useful methodology to determine the accuracy of reliability predictions and assessments. The results of this study indicate that the following techniques if properly developed (i.e. user friendly) may be the answer to the problem:

1) Expert systems
2) Reliability Analyzer.

The development of expert systems for carrying out various phases of system reliability analysis is, as of today, in its infancy but likely to pick up vary fast. Hollick (Ref 33) used the Fuzzy Set Theory to develop an expert system, while Gordon and Shortcliffe (Ref 34) used the Dempster Shafer Theory for their MYCIN project. It appears that there is a tremendous possibility of developing this area of high technology into practical tools for the assessment of reliability of various systems.

In addition to the new software technology described above, there is also a hardware available called Reliability Analyzer, which can be used as a user friendly practical estimator of system reliability. The concept of this new hardware device was originally presented by Misra (Ref 35) and was developed further by Bansal and Jain

(Ref 36,37). Laviron, Carino and Manaranche (Ref 38) used this concept to produce a commercial Reliability Analyzer called ESCAF (Electronic Simulator for Computing and Analyzing Failures). A more versatile and user friendly version called S.ESCAF was later developed by Blot and Laviron (Ref 39). The results of this study show that the current developments in Expert System and Reliability Analyzer will lead to many forms of commercially available practical reliability assessment gadgets in the near future. Additionally these Expert Systems and Reliability Analyzers should have built in precision techniques for determining the accuracy of the reliability predictions.

CONCLUSIONS

The Summer Research project designed as a complementary effort to Rome Lab's on going "New System Reliability Assessment Methods" program has been completed within the specified time of twelve weeks. This preliminary research project has clearly indicated the future research work necessary in this field.

RECOMMENDATIONS FOR FOLLOW ON RESEARCH EFFORTS

It is recommended that the following research tasks should be pursued as further complementary efforts to the "New System Reliability Assessment Methods program":

1. A simulation model should be developed to estimate the confidence intervals for the capacitor data identified in Task 2. The model should be developed further from the single component level to the multicomponent (circuit board) level, paving the path for further development of the model to the system level.

2. A detailed study of Fuzzy Sets Theory and Dempster-Shafer Evidence Theory should be undertaken for applications in the field of reliability assessments and their accuracy for electronic systems.

3. The role of Expert Systems and Reliability analyzers as possible user friendly practical devices for the estimation of reliability should be further investigated.

ACKNOWLEDGMENT

The Author would like to thank Joe Caroli, the Focal point of the project and Jim Collins, Chief, ERSR Branch Rome Lab, Rome N.Y. for their encouragement, technical help and hospitality during this summer project. Thanks are also extended to Mrs Josie Mirowski for typing this report.

REFERENCES

1. Denson, William and Keene, Samuel, "New System Reliability Assessment Methods." RAC Project #A06830, January 1996.

2. Denson, William and Keene, Samuel, "A New Systems Reliability Assessment Method." Reliability Review, Vol 15, p 16 - 20, December 1995.

3. Judnick, W. E., "Confidence Intervals for System Failure Rates: A Literature Review." 1985 IEEE Proceedings Annual Reliability and Maintainability Symposium.

4. Miller, P. E. and Moore, R. I., "Field Reliability Versus Predicted Reliability: An Analysis of Root Causes for the Difference." 1991 IEEE Proceedings Annual Reliability and Maintainability Symposium.

5. Harris, N. and O'Connor, P. D. T., "Reliability Prediction: Improving the Crystal Ball." 1984 IEEE Proceedings Annual Reliability and Maintainability Symposium.

6. Tomsky, J., "Analysis of Variance of Reliabilities." 1983 IEEE Proceedings Annual Reliability and Maintainability Symposium.

7. Nachlas, J. A., Guber, S. S., and Wiesel, H. Z., "Sensitivity in Weibull System Reliability Models." 1984 IEEE Proceedings Annual Reliability and Maintainability Symposium.

8. Huang, Z. and Porter A. A., "Lower Bound on Reliability for Weibull Distribution When Shape Parameter is not Estimated Accurately." 1991 IEEE Proceedings Annual Reliability and Maintainability Symposium.

9. Spencer, J. L., "The Highs and Lows of Reliability Predictions." 1986 IEEE Proceedings Annual Reliability and Maintainability Symposium.

10. Prairie, R. R. and Zimmer, W. J., "An Iterative Bayes Procedure for Reliability Assessment." 1990 IEEE Proceedings Annual Reliability and Maintainability Symposium.

11. Romeu, J. L., "Confidence Bounds for System Reliability." RAC Report #SOAR-4, Spring 1985.

12. Denson, W. K., "Reliability Assessment of Critical Electronic Components." IIT Research Institute Report RL -TR-92-197, July 1992.

13. Military Handbook - Reliability Prediction of Electronic Equipment, MIL-HDBK-217F, February 1995.

14. Denson, W. K. and Priore, M. G., "Automotive Electronic Reliability Prediction." International Congress and Exposition, Detroit, Michigan, February 1987, SAE Report #870050.

15. Devone, J. L., "Probability and Statistics for Engineering and the Sciences." Duxbury Press ITP, 1995.

16. Amstadler, B. L., "Reliability Mathematics - Fundamentals, Practices, Procedures." McGraw-Hill Book Company.

17. Aggarwal, K. K., "Reliability Engineering." Kluwer Academic Publishers, 1993.

18. Zellen, I. M. (editor), "Statistical Theory of Reliability." Mathematics Research Center, US Army, The University of Wisconsin, 1964, p 115 - 137.

19. Sprinter, M. D. and Thompson, W. E., "Bayesian Confidence Limits for the Product of N Binomial Parameters." Biometrika, Vol 53, 1966, p 611-613.

20. Levy, L. L. and Moore, A. H., "A Monte Carlo Technique for Obtaining System Reliability Confidence Limits From Component Test Data." IEEE Transactions on Reliability, Vol R-16, No. 2, September 1967, p 69 - 72.

21. Sandquist, G. M., "Introduction to System Science." Prentice-Hall, Inc.

22. Faurre, P. and Depeyrot, M., "Elements of System Theory." North-Holland Publishing Company.

23. Zadeh, L. A. and Polak, E., "System Theory." McGraw-Hill Book Company.

24. Gleick, J., "Chaos - Making of a New Science." Penguin Books, Inc.

25. Misra, K. B. (editor), "New Trends in system Reliability Evaluation." Elsevier Science Publishers, 1993.

26. Zadeh, L. A., "Fuzzy Sets." Information and Control, Vol 8, p 338 - 353, 1965.

27. Zadeh, L. A., "Fuzzy Sets as a Basis for a Theory of Possibility." Fuzzy Sets and Systems, Vol 1, No. 1, p 3 - 28, 1978.

28. Dempster, A. P., "Upper and Lower Probabilities Induced by a Multi-Valued Mapping." Ann. Math. Statist., Vol 38, p 325 - 339, 1967.

29. Shafer, G., "A Mathematical Theory of Evidence." Princeton University Press, Princeton, New Jersey, 1976.

30. Misra, K. B. and Sharma, A., "Performance Index to Quantify Reliability Using Fuzzy Subset Theory." Microelectronics and Reliability, Vol 21, No. 4, p 543 - 549, 1981.

31. Keller, A. Z. and Kara Zaitri, C., "Further Applications of Fuzzy Logic to Reliability Assessment and Safety Analysis." Microelectronics and Reliability, Vol 29, No. 3, p 399 - 404, 1989.

32. "Expanded Confidence Assessment Process (ExCap) Application to PAC-3 Missile Reliability." Nichols Research, Inc. US Army Missile Command, Redstone Arsenal, AL 35898, March 1996.

33. Hollick A., "DIAFUZZY (VERSION 1): An Inference - Engine for Approximate Reasoning." Interatom GmbH, Berglisch Gladbach, Germany, September 1987.

34. Gordon J. and Shortcliffe, E. H., "Dempster - Shafer Theory of Evidence and Its Relevance to the Expert System, In Rule-Based Expert Systems - The MYCIN Experiments of the Stanford Heuristic Programming Project." Chapter 13, ed. Buchanan, B. G. and Shortcliffe, E. F., Addison-Wesley, Reading, 1984.

35. Misra, R. B. and Raja, A. K., "A Laboratory Model of System Reliability Analyzer." Microelectronics and Reliability, Vol 19, No. 3, p 259 - 264, 1979.

36. Bansal, V. K. and Misra, K. B., "Hardware Approach for Generating Spanning Trees in Reliability Studies." Microelectronics and Reliability, Vol 21, No. 2, p 243 - 253, 1981.

37. Bansal, V. K., Misra, K. B., and Jain, M. P., "Minimal Pathsets and Minimal Cutsets Using a Search Technique." Microelectronics and Reliabiity, Vol 22, No. 6, p 1067 - 1075, 1982.

38. Laviron A., Camino, A., and Manaranche, J. C., "ESCAF - A New and Cheap System for Complex Reliability Analysis and Computation." IEEE Trans. Rel., Vol R-31, No. 4, p 339 - 349, October 1982.

39. Blot, M. and Laviron, A., "Reliability Analysis With the Simulator S.ESCAF of a Very Complex Sequential System: The Electrical Power Supply System of a Nuclear Reactor." Rel. Engrg. and Syst. Safety, Vol 21, No. 2, p 91 - 106, 1988.

40. Pipes, L.A. and Harvill, L.R. "Applied Mathematics for Engineers and Physicists". p 484-486; p543-546. McGraw-Hill Book Company.

THE ANALYSIS OF *PROFILER* FOR MODELING THE DIFFUSION OF ALUMINUM-COPPER ON A SILICON SUBSTRATE

Matthew E. Edwards
Associate Professor Of Physics
Department of Physics

Spelman College
350 Spelman Lane
Atlanta, Ga. 30314-4399

Final Report For:
Summer Faculty Research Program
Rome Laboratory

Sponsored By:
Air Force Office of Scientific Research
Bolling AFB, Washington, D. C.

September 1996

THE ANALYSIS OF *PROFILER* FOR MODELING THE DIFFUSION OF ALUMINUM-COPPER ON A SILICON SUBSTRATE

Matthew Edwards
Associate Professor of Physics
Department of Physics
Spelman College

Abstract

A detailed analysis of ***PROFILER***, a computer program that predicts the interdiffusion of an alloying couple, has been completed, and its application to the diffusion of copper into aluminum has been initiated. ***PROFILER*** has been observed to model the concentration profiles of up to eight diffusants in the alloying couple. Also, the program provided its results quickly, within a matter of seconds. While its applicability to Aluminum-Copper thin films remains to be established, such as effort within itself would be viable and should be completed. These considerations, on diffusion, were a part of a larger effort, the systematic prevention of electromigration, which occured from electrical current-induced movement (diffusion) of metallic atoms. The analysis of ***PROFILER,*** as provided by thermal stressing of the sample, and its preliminary application to an Aluminum-Copper thin film on a silicon substrate are reported as an effort to understand and control electromigration.

THE ANALYSIS OF *PROFILER* FOR MODELING THE DIFFUSION OF ALUMINUM-COPPER ON A SILICON SUBSTRATE

Introduction

Interdiffusion is an important process in reliability physics. It is the process that is occurring when the components in connecting alloys diffuse across the plane of coupling. This process finds itself applicable in issues associated with interconnects, composite materials, high-temperature coatings, and thin-film devices [2-6]. Its importance for the reliability of interconnects is in the prevention or at the very minimum the prediction of electromigration, where the latter is an undesirable phenomenon of atomic migration. The region of atomic migration is referred to as the diffusion layer. This layer is often the region where corrosion, electrical anomalies, embrittlement and other deleterious processes occur. In the cases of high-temperature coatings and thin-film devices, for instance, interdiffusion can lead to early electronic device failures. Interdiffusion, as brought on by electromigration from currents in circuits, can also lead to device failure. Therefore, there is a need to better understand the nature of metallic diffusion, and to have predictive analyses of the diffusants' concentrations as a function of time and penetration depth.

Therefore, the objective of this investigation has been to evaluate the computer model program, ***PROFILER*** [1], as it relates to atomic migration . The program, using the square-root diffusivity method [6] as a solution to Fick's second law (the time dependent diffusion equation), generates the concentration profiles of the diffusants of the system. In addressing this objective, the essential question is how well does ***PROFILER*** predicts the interdiffusion of Cu-Al components following thermal stressing. Although the answer re-

mains underdetermined from this preliminary effort, significant progress has been made and the results are presented here. ***PROFILER*** has been shown to be excellent at modeling concentration profiles, and its salient features are presented in the following sections.

Discussion

The operations of ***PROFILER*** (from the DOS prompt or through windows) are implemented through a pull-down menu, as described below. Each item will be described to varying degrees, as provided or determined from this preliminary consideration (See Table 1. for a layout of the main menu):

1. Load - This sub-menu item reactivates existing or previously prepared files into the operating program. Pressing <Enter> displays on the screen all file names that have been saved in prior sessions. Select the desired file and press <Enter>.

Note: The Load entry should not be selected until files have been appropriately saved. If selected before files are saved, the program gives an error message.
(The user should start with the New screen, as described later, to avoid this problem).

2. Save - This item saves the current data. The program prompts the user to enter a file name having up to 8 letters or numbers. In all cases, saved information items or selections to ***PROFILER*** are implemented by pressing <F2>.

Table 1: ***PROFILER'S*** Main Menu

Files	Data	*Graphics	*Numeric
1. Load	**7. System Information**	**15. **Diffusion Time**	**18. Diffusion Time**
2. Save	**8. ****L matrix**	**16. Concentration Differences**	**19. Concentration Differences**
3. Save As	**8.1 Frequency Factors**	**17. ***Display Concentration Profiles**	**20. ***Vector Matrix A**
4. New	**8.2 Activation Energies**		**21. ***Generate Numeric Data – Commas**
5. Exit (save first)	**8.3 Tracer Diffusivities**		**22. ***Generate Numeric Data – Spaces**
6. Quit (no save)	**8.4 L matrix**		
	9. ****G matrix		
	9.1 Regular Solution Parameters		
	9.2 G matrix		
	1. D matrix		
	11. r matrix		
	12. Eigenvalues of r		
	13. Eigenvectors of r		
	14. Alpha matrix		

*Selective only after System Information fill-in sheet and the D matrix have been completely implemented.
**A diffusion time always exist, having a default value of 360,000 seconds.
***Further selective only after concentration Differences have been specified.
****Has second level menu items as shown

3. Save As - This item allows the user to change the name of an existing file that is currently running. The existing file name appears after Save As on the main menu screen.

4. New - This item is used to create a new file. It should be the first selection made for a beginning session with ***PROFILER***. When selected, a new System Information screen is displayed and the user must enter the following information about the diffusion couple:

   a. The number of alloying elements (diffusants) in the couple. The program has a default value of 2.

   b. The temperature (in Kelvin) at which diffusion takes place. The program has a default value of 1500K.

   c. Average mole fractions of the solutes in the couple. Note: The program calculates the average mole fraction of the host element, from information about the solutes.

   d. Abbreviations for each alloying element (optional).

   e. MO (Structural Factor) -This item has a default value of zero. It gives information on the diffusion steps which may go backwards rather than forward. Typical values are as follows:

Simple cubic	- 3.77	Body Centered Cubic	- 5.33
Face Centered Cubic	- 7.15	Diamond	- 2.00.

If the appropriate information is not entered on the System Information Screen, the pro-

gram gives an error message and automatically stops running.

5. Exit (save first) - This item will exit the program and save the data under the user's specified filename.

6. Quit (no save) - This item will exit the program without saving the entered data. The program has a safety feature which prompts the user to save the data before exiting the program.

7. System Information - This item is the same as that available under # 4 above. It gives pertinent information about the current diffusion couple, and can be selected at any time while using the program.

8. L matrix - This item is where tracer diffusion data are entered.

8.1 Frequency factor (A) - (Not considered in this preliminary investigation. The program runs and gives satisfactory results without specifying the frequency factor.)

8.2 Activation energies (Q) - The typical energies for migration of elements.

8.3 Tracer diffusivities (D). - The typical diffusion coefficients of elements.

8.4 L matrix - (Not considered in this preliminary investigation. The program runs and gives satisfactory results without specifying the L matrix.)

9. G matrix - This is a pull-down menu to enter thermodynamic information about the diffusants or elements.

9.1 Regular Solution Parameters - These values can be approximately related to heats of mixing, ΔH^{mix}, mole fractions, N_i, and other measured thermodynamic quantities by various formulas, e.g.:

$$\omega_{ij} \cong \frac{\Delta H^{mix}}{N_i N_j}.$$

9.2 G matrix - The second derivative of the free energy are entered here in units of J/mole, where

$$G_{ij} = \frac{\partial^2 G}{\partial N_i \partial N_j}.$$

10. D matrix - The diffusivity matrix elements [6-8] are entered here if they are known.

11. r matrix - The square root diffusivity matrix [r] is entered here.

Note: If L and G matrices are provided, then ***PROFILER*** will automatically calculate D and r. If D or r is provided, then program will calculate the one that's not entered. The calculated diffusivities appear on their respective menu screens.

12. Eigenvalues of r - The square roots of the eigenvalues of D are displayed (i.e. $r_i = \sqrt{D_i}$).

13. Eigenvectors of r - The α^{-1} matrix is displayed. Columns of this matrix are eigenvectors of both D and r. The α and α^{-1} matrices diagonalize D by the transformation $D_t = \alpha D \alpha^{-1}$.

14. Alpha matrix - This item displays the α matrix.

15. Diffusion Time - The isothermal heat treatment time is entered in units of seconds. The program has a default time of 360,000 seconds (100 hours).

16. Concentration Differences - The initial concentration differences between the diffu sion couple are entered. The values are obtained by subtracting concentrations on the left side of the coupling interface from those on the right side.

17. Displays Concentration Profile - The monitor's screen shows the concentration profiles of each solute. The axes variable are the followings:

X axis - Gives the distance from the initial interface (a distance of 10 μm is between graduations marks on the X axis).

Y axis - Gives the difference in concentration between the local concentration and the average concentration for each solute. (0.2 of units of the concentration are between graduations on the Y axis).

The purpose of the Numeric menu is to create a file of concentration profiles in a tabular form. This file can then be imported into a spread sheet or plotting program to obtain a hard copy [9] of the concentration profiles or to construct diffusion paths.

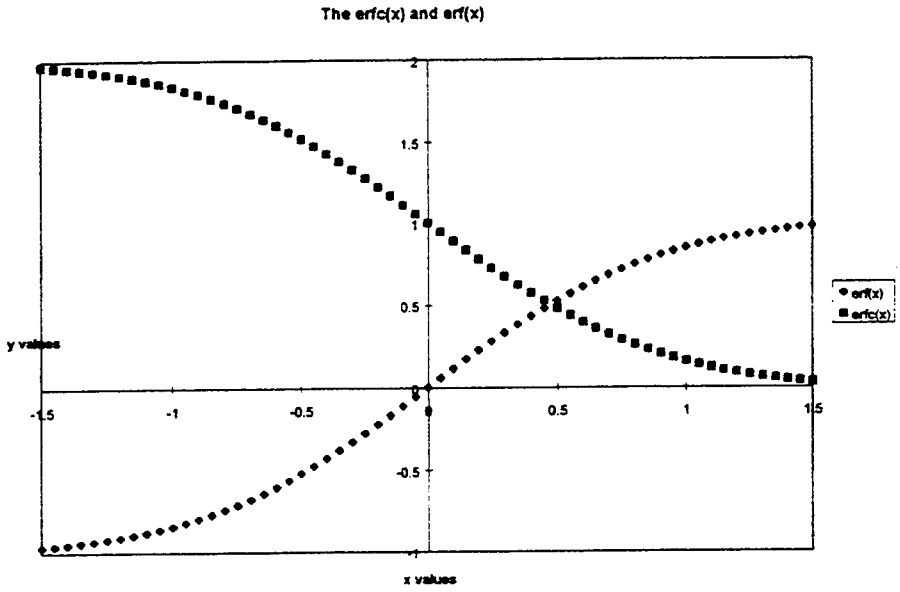
18. Diffusion Time - Displays the time as entered in # 15 above.

19. Concentration Differences - Displays the initial concentration differences as used to calculate the concentration profiles.

20. View matrix A - The concentration profile for an n component diffusion couple is given by the sum

$$C_i(x,t) = C_i^R - \sum_{i=1}^{n-1} A_{ij} erfc\left(x / (2\sqrt{D_i t})\right),$$

where C_i^R is the initial concentration of solute i in the alloy on the right side of the initial interface, and D_j is one of the n-1 eigenvalues of the D matrix [5-6]. The D matrix gives the presentation of the A_{ij} coefficients of the erfc function, where the latter is a special mathematical function [15]. Its graphical representation is as shown adjacently.



21. Generate Numeric Data -- Commas - This item creates a file with numeric data separated by commas.

22. Generate Numeric Data -- Spaces - This item creates a file with numeric data separated by spaces.

The format of the output file is the following:

column 1	column 2	column 3	column n.
-0.00100	0.2	0.3	-1.4

Column 1 gives x axis values in centimeters (x = 0 is the original-finite interface between the alloying couple).

Column 2 gives concentration differences for the first solute between the local . concentration and the average initial concentration , where

$$\Delta C_i = C_i(x,t) - \frac{C_i^R + C_i^L}{2}.$$

(The concentration difference is always zero at x = 0).

Column 3 gives the concentration differences for the second solute.

Column n gives the concentration differences for the n-1 solute.

Next, a systematic usage is made of these menu items to implement ***PROFILER*** under particular conditions.

Methodology

The outcome from ***PROFILER*** can been obtained from one of two different modes of operation, depending on the type and amount of information that's initially known about the diffusing couples [11-17]. Table 2, below, outlines the two different modes of operation:

USING ***PROFILER*** TABLE 2

Method I.	**Method II**
Pre-Determined Diffusivity Method 1. System Information a. Number of Components in the interdiffusion-couple. n has the value: (1 < n < 8), and must be the first entry. b. Temperature in Kelvin at which the diffusion process occurs. c. Average Mole fraction (at. Pct.) of each solute between the left member and the right member of the interdiffusion couple. (At this point the program calculates the average fraction of the solvent, the "host" element). d. Enters abbreviations of alloying element. (optional entry - but helpful for	Undetermined Diffusivity Method (Same as Method I for a - d.)

accountability)

e. Structure Factor (MO) Unnecessary Entry for this method. (The appropriate value has already been considered in the Pre-determination of the diffusivity)	e. Structure Factor (MO) As determined from diffusion theory: 1. simple cubic - 3.77 2. body centered cubic - 5.33 3. Face centered cubic - 7.15 4. Diamond - 2.00
2. Diffusivities Not entered.	2. a. Tracer diffusion Values (Highlight L matrix on menu bar). Pre-exponential (A) units of cm^2/sec, Activation Energies (Q) units of kcal/mole b. Thermodynamic Data - Heat of mixing delta H^mix, Mole fraction of N_i and N_j of solutes
3. Diffusion Time in seconds	3. (Same as method I)
4. Concentration Differences Enter the concentration of right member of the interdiffusion couple minus the concentration of the left member.	4. (Same as method I)
5. Displays Concentration differences (difference between the absolute concentration minus the average as entered for the diffusion couple): Ordinate axis, in units of at. Percent Abscissa axis, in units of cm.	5. (Same as method I.)

Results

Preliminary results are shown in Fig. 1 - Fig. 3. In Fig. 1 the raw data, as produced from the Auger Spectrometer, are reproduced in their natural representation of concentrations as a function of sputter time. All experimental data are provided in this format. The graphs of Fig. 1 show the experimentally determined migration from thermal stressing of Carbon C, Oxygen (O), Copper (Cu), Aluminum (Al), and Silicon (Si). However, for these con-

centrations to be compared with the modeling of ***PROFILER***, the sputter time in minutes must be changed to nanometers. This has been done as shown in fig. 2 using a sputter rate of 6.3 nm/min, the sputter rate of Aluminum. Fig. 3 is an actual output of ***PROFILER*** for the simple arrangement of pure copper initially diffusing into initially pure aluminum and vice versus. The output, in fig. 3, is for a finite-interface diffusion process. The ideal arrangement, of a finite-interface, needs to be adjusted in consideration of the operating conditions of the Auger Spectrometer and the corresponding samples.

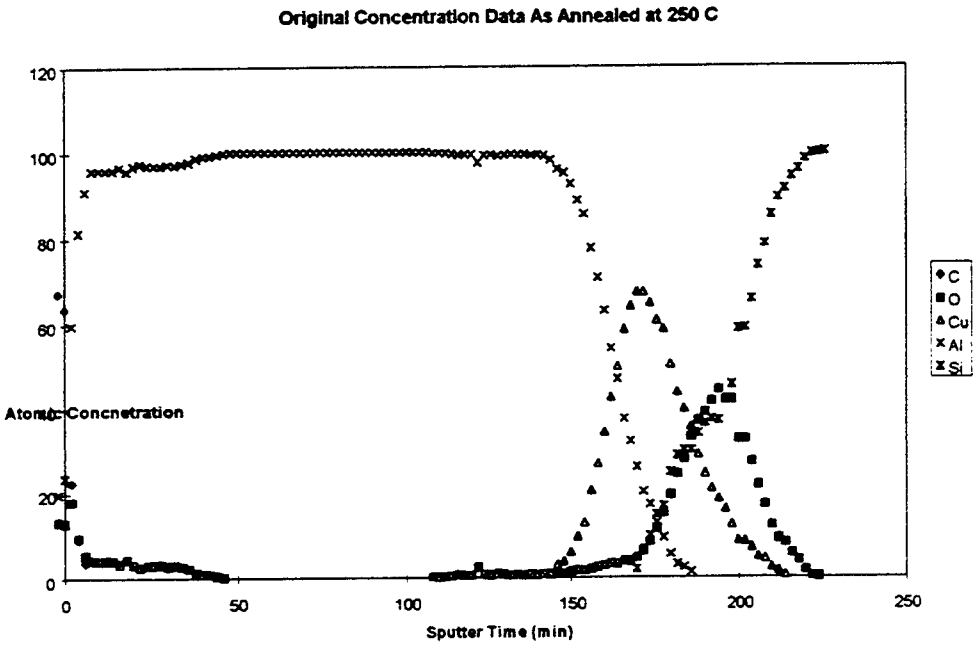


Fig. 1. Shows the Concentration Profiles As Produced By the Auger Spectrometer.

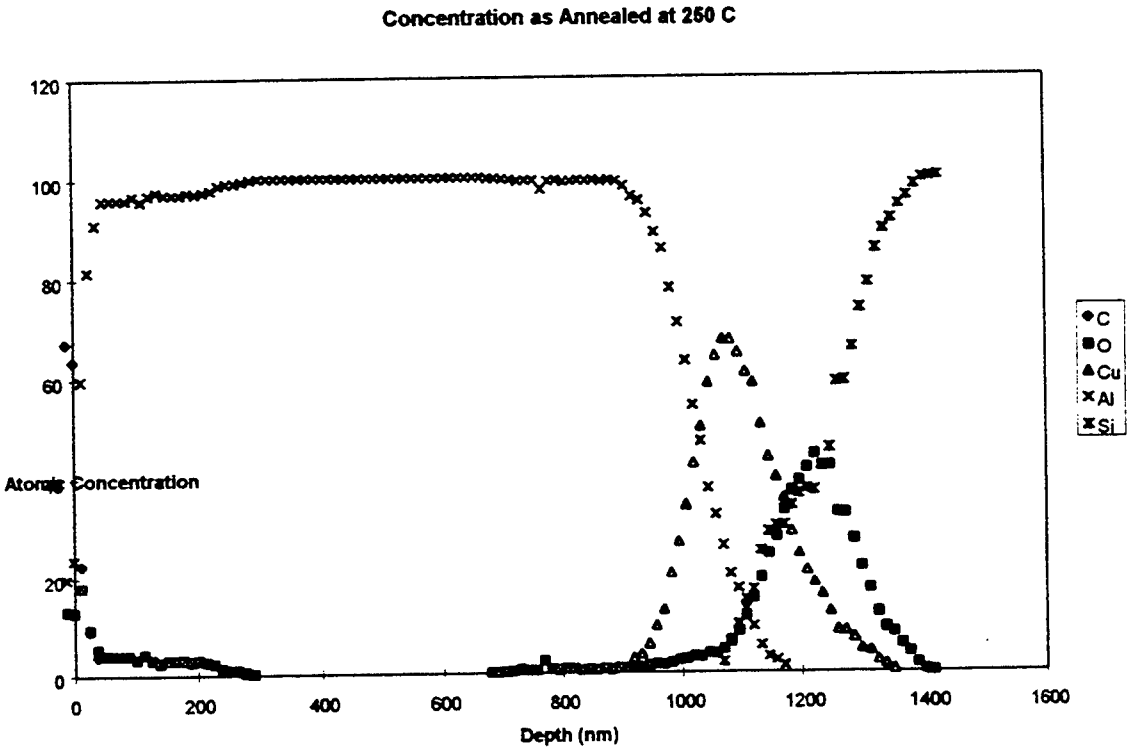


Fig. 2. Shows the Concentration Profiles as a Function of Depth into the Sample.

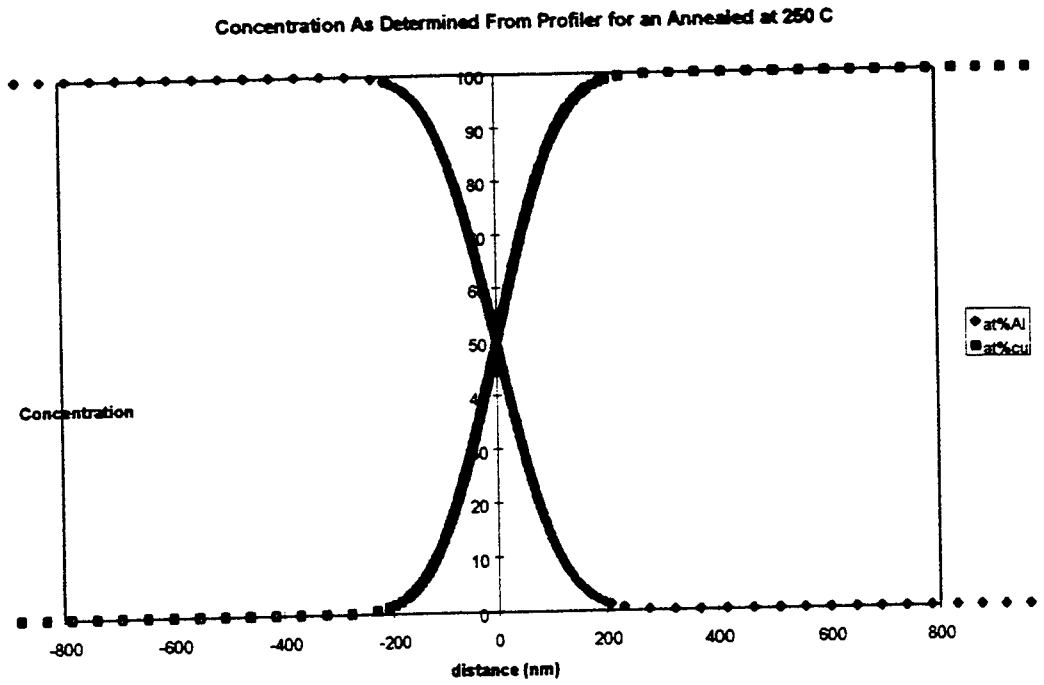


Fig. 3. Concentration Output of ***PROFILER***.

Conclusion

The investigator has found ***PROFILER*** to be a powerful and efficient program for modeling the diffusion of metallic atoms across a finite initial plane of coupling for two alloying components. It produces the concentration profiles in a matter of a few seconds for up to eight diffusants in the alloying couple. However, its applicability to a dispersed (variable) initial interface remains undetermined from this preliminary effort. In fact, the following additional issues must be considered before a determination is made on the applicability of ***PROFILER*** to the Al-Cu samples of interest: 1. How reliable are the results as provided by ***PROFILER*** to samples having elements as opposed to alloys in the initial couple? and 2. How might the dispersion of an initial interface be accounted for programmatically either before or after the application of ***PROFILER***? Once these issues are clearly resolved, then the applicability of ***PROFILER*** to specific Al-Cu samples having these attributes will become a certainty.

References

1. ***PROFILER:*** *Diffusion Couple Software to Predict Concentration Profiles and [D],* as written by William B. Brockman, at the University of Connecticut, and Available from Johhn E. Morral, Dept. of Metallurgy and Institute off Material Science, University of Connecticut, tel: 203 486-2923, FAX 203 486-4745

2. *Elements of Material Science*, 2 nd. Ed. Lawrence H. Van Vlack, Addison-Wesley Publishing Co., 1964.

3. *Copper Diffusion Into Aluminum-Silicon Metallization By Accelerated Thermal and Electrical Stressing,* G.O. Ramseyer, L.H. Walsh, J.V. Beasock, H.F. Helbig, R..C. Lacoe, and S. Brown, in press.

4. *Handbook of Auger Electron Spectroscopy*, 2 nd. Ed., Lawrence E. Davis, et. Al., Physical Electronics Division, Perkin-Elmer Corporation, 1976.

5. *Classification Of Concentration Profiles In Quaternary Diffusion Couples*, M.K. Stalker, and J.E. Morral, Acta Metall. Mater. Vol 38, No. 3, pp 439-447, 1990.

6. *Applications of the Square Root Diffusivity to Diffusion in Ni-Al-Cr Alloys*, M.S. Thompson, J.E. Morral, and A.D. Romig, Jr., Metallurgical Transactions A, Vol 21A, Oct. 1990, p 2679.

7. *MULTICOMPONENT DIFFUSION: Implementation of the Square-Root Diffusivity Method Via the* ***PROFILER*** *Computer Program.* T.H. Cohen and M.E. Glicksman, Modelling Simul. Mater. Sci. Eng. **3**, (1995) 585-596.

8. *Interdiffusion in the Ni-Cr-Co-Mo System at 1300 C*, J.A. Heaney, III and M.A. Dayananda, Metallurgical Transactions A, Vol 17A, June 1986, pp 983
9. Microsoft Manuals on Microsoft Windows95, Word 7.0, Excel 7.0, Equation Editor, etc.
10. Private Communications with Professor Morral, University of Connecticut

11. *Thin Film Phenomena*, Kasturi L. Chopra, McGraw-Hill, 1969.

12. *Elements of X-Ray Diffraction*, B.D. Cullity, Addison-Wesley Publishing Co, 1978.

13. Characterization off Copper Diffusion Into Al and Al-1% Si Polycrystalline Thin Films, L.H. Walsh, G.O. Ramseyer, J.V. Beasock, H.F. Helbig, In Press.

14. *Handbook Of Chemistry and Physics*, 6 th. Ed., CRC Press, 1985-86.

15. *Advanced Mathematics For Engineers and Scientists,* Murray R. Spiegel, Schaum's Outline Series, McGraw-Hill, 1993.

16. *Phase Transformation in Metals and Alloys*, D.A. Porter and K.E. Eastering, Van Nostrand Reinhold, 1982.

17. *Selected Values of the Thermodynamic Properties off Binary Alloys,* Ralph Hultgren, Pramod Desai, and Donald Hawkins, American Society For Metal, 1973.

SPEAKER IDENTIFICATION AND ANALYSIS OF STRESSED SPEECH

Kaliappan Gopalan
Associate Professor
Department of Engineering

Purdue University Calumet
Hammond, IN 46323

Final Report for
Summer Faculty Research Program
Rome Laboratory

Sponsored by
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

July 1996

SPEAKER IDENTIFICATION AND ANALYSIS OF STRESSED SPEECH

Kaliappan Gopalan
Associate Professor
Department of Engineering
Purdue University Calumet
Hammond, IN 46323

Abstract

In the first part of this project, feature extraction using Fourier and Fourier-Bessel (FB) transforms was carried out for the purpose of text-independent speaker identification. It was found that for speech transmissions from aircraft, a combination of 20 cepstral coeffcients on linear and mel frequency scale yielded an identification score of 80 %. A slightly lower score of 76 % resulted when the features were formed using log spectral energies in 20 bands of overlapping frequencies. Identification scores of 74 % and 76 % were achieved using a set of 15 and 20 features based on the expansion of the speech signals in the FB transform. Due to the highly noisy nature and the short segments of the test data base, feature vectors obtained from the linear predictive representation of speech, however, yielded poor identification scores of below 55 %. The scores in each case were obtained with a single set of features using the same commercial classifier that was based on vector quantization of features. The single-feature based results achieved in this project compare favorably with the results obtained on the same speech data base using methods of feature and/or classifier fusion at Rome Laboratory.

The same set of features based on Fourier and FB transforms were studied for the identification of speakers using a second group of nine speakers. The utterances for this group consisted of aircraft-to- ground transmissions of speech by nine pilots who were considered under stress. With 1054 test utterances, scores of 88 % and 84 % resulted using 20 cepstral coefficients and 20 log spectral energies respectively. Using FB transform-based features, the scores achieved were 65 % with the energy parameters and 62 % with the frame difference of the energy parameters.

Based on the identification scores using cepstral and FB transforms, a study of the analysis of speech under stressed conditions was begun in the second part this project. The initial results using the FB transform appear to show variations of features with variations in the stress level of the speaker under mayday conditions. Further processing using FB transforms is expected to better bring out the acoustical correlates of speech under stress.

Speaker Identification and Analysis of Stressed Speech

Kaliappan Gopalan

I. Introduction

Identification of speakers based on their speech signals has many commercial and military applications. In particular, automatic recognition of pilots based on their voice transmissions from the aircraft is important for fast and accurate strategic decision-making under adverse conditions.

In general, an automatic speaker identification system consists of two major blocks, namely, a feature extractor and a classifier. The feature extractor determines a set of compact and efficient temporal parameters (a reference feature set) that represents the time-domain utterance of a speaker in a chosen feature domain. During the training phase of the system, one or more sets of features are stored in a library for each member in the group of speakers to be recognized. A feature set arising from an utterance of a speaker during the testing or identifying phase is compared by the classifier against each stored set of reference features. By using a minimum distance criterion, the classifier outputs the identity of the speaker whose reference feature set is closest to the unknown feature set as the most likely source of the test utterance.

Although some of the common issues that plague commercial identification systems such as reluctant speech and voice disgusing are generally not present in a system for military applications, other more serious problems arise due to speech transmissions from fighter aircraft. The engine noise, which varies with the altitude of the aircraft, and the breath noise of the pilot, for example, contribute significantly to the received noise power and contaminate the speech. In addition, the transmission channel, the method of transmission and the bandwidth of the receiver adversely affect the integrity of the received signal. Consequently, the temporal patterns obtained from the transmissions do not accurately represent the speech or the speaker. A short duration of speech during the testing phase further complicates the correct identification since all of the characteristics of the speaker may not be present within the available duration. All of these problems lead to errors in the identification of speakers even with the best possible classifier. An efficient parameter set that captures the characteristics of a speaker while rejecting all extraneous information present in the received signal greatly alleviates the identification problem.

One of the goals of the summer research was to investigate the efficacy of the commonly used spectral domain features and the proposed Fourier-Bessel transform-based features. The different sets of features considered were tested on a speech data base that has utterances from 41 speakers (pilots). The data base, known as the Greenflag data base, consisted of digitized speech transmissions of the speakers from aircraft. For comparison of the efficiency of the identification system, the same classifier was used with each of the feature sets chosen.

Apart from speaker identification, a feature set may yield information that relates to the stress level of the speaker. This information is useful in determining the workplace stress level, and in monitoring the physiological and emotional state of an aircraft pilot under adverse conditions. Additionally, the study of the variation of speech parameters under stress may help toward the design of better speech and speaker recognition systems. As the second goal of the summer research, a preliminary study of the acoustic correlates of speech under stress was to be carried out using two data bases. The first data base consisted of speech transmissions from a group of nine European speakers (fighter pilots) and the other included transmissions from fighter pilots under a mayday (flameout) condition.

For the study of the parametric variations in stressed speech, spectral energy in the vicinity of fundamental frequency and at high frequency range were observed both in the Fourier and FB domains.

The following sections describe the features studied, the observed speaker identification results for the Greenflag and the stressed speech data bases, and the preliminary results observed from the stressed speech processing.

II. Feature Extraction

For all the parameters discussed below, a frame length of 25 ms (200 samples at the sampling rate of 8000/s) was used with an overlap of 12.5 ms. Initially, to verify the commercial classifier performance, the following parameters were extracted from a 19th order linear prediction (LP) model to form feature vectors for each frame of speech.

(a) 19 predictor coefficients and the prediction error
(b) 19 reflection coefficients and the prediction error

Features from a set of transmissions from the Greenflag data base were used to form a vector-quantized (VQ) codebook using a software package by Entropic Research Laboratory. The VQ classifier from the same package, however, yielded identification scores of below 60 % when tested with the above features for 143 test transmissions. Although the LP model is most commonly used to represent the speech production mechanism, it performs poorly for signals under noisy conditions and for nonvowel sounds [1], as evidenced by the low scores of the overall system. Therefore, LP-based parameters were not pursued further as a single feature set in this study.

The second set of parameters considered were a set of cepstral coefficients based on the mel scale. Initially, 19 cepstral coefficients and the log energy of each speech frame at the selected cepstral indices were evaluated as follows [1,2]. With a 1024-point discrete Fourier transform (DFT) of each frame, the frequency spectrum was obtained at a resolution of approximately 8 Hz/point. Starting at approximately 100 Hz (13 in DFT index) - the fundamental frequency for

the all-male data base - 10 frequencies were chosen on a linear scale with a spacing of 100 Hz. Above 1000 Hz, mel frequency scale was used to cover the range from 1100 Hz to 3500 Hz. Outputs of filters centered at the chosen frequencies were formed from the spectrum of each frame as shown in Fig. 1. The filters for the linearly distributed range of frequencies had a constant bandwidth of 100 Hz while those for the mel scale (critical band filters) had logarithmically increasing bandwidths with each filter spectrum overlapping the adjacent filter spectra. The choice for this was motivated by the observation that the auditory system perceives information based on the energy in a band of frequencies rather than that at a single frequency and that the band increases with the increase in frequency.

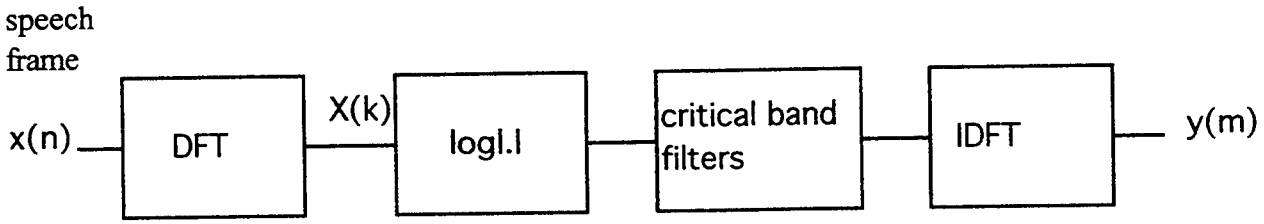


Fig. 1 Computation of the cepstral coefficients

The 19 cepstral coefficients were obtained as the inverse DFT of the log energy at the selected indices The sum of the filter responses, which represents the total log energy, formed the 20th feature element.

To represent the spectral changes from frame to frame, the delta cepstrum was also computed from the 20-element cepstral feature for (a) three frame difference and (b) five frame difference.

Nonsinusoidal representation

The other set of features considered to represent speakers was obtained from the Bessel function representation of the speech signals. Inasmuch as speech , particularly voiced speech, has quasiperiodic variation of amplitude with time, a set of basis functions that has a similar quasiperiodic behavior is more suitable to represent speech than the periodic sinusoidal set. Bessel functions possess the quasiperiodic amplitude variation with gradually decaying amplitude, resembling the behavior of voiced speech within a pitch period. Therefore, the resulting time-domain representation of speech using an orthogonal set of Bessel functions is efficient in retaining the speech quality [3, 4]. The Bessel function representation was also used successfully on a small data base for discrete utterance and text-dependent speaker identification [5]. These previous results motivated the representation of speakers for identification using the Bessel function $J_1(t)$.

Fourier-Bessel Representation

An arbitrary function x(t) in the interval $0 < t < a$ is represented using $J_1(t)$ in the Fourier-Bessel expansion given by

$$x(t) = \sum_{m=0}^{\infty} C_m J_1([x_m/a]t), \; m = 1, 2, 3, \ldots \quad (1)$$

where x_m, $m = 1, 2, 3, ..$ are the roots of $J_1(t) = 0$.

Using the orthogonality of the set $\{ J_1([x_m/a]t) \}$, the coefficients of expansion are determined from

$$C_m = (2/[a^2.J_0(x_m)]) \int_0^a t.x(t).J_1([x_m/a]t)dt \quad (2)$$

As with the Fourier series, the above Fourier-Bessel coefficients $\{ C_m \}$, are unique for a given x(t). Unlike the sinusoidal basis functions in the Fourier series, however, the Bessel functions are aperiodic and decay within the range a.

Speech Signal Representation and Feature Extraction

As with the spectral domain description of speech, representation using the Bessel function set is carried out on short intervals of speech to preserve the time-varying nature . For the Greenflag data base, which was obtained at the sampling rate of 8000/s, the same interval of 25 ms (200 samples/frame), as with the cepstral domain description, was used for the range a. With a frame overlap of 12.5 ms, each frame of speech can be represented by the FB coefficient set, $\{ C_m \}$.

Although the summation in Eq. (1) goes to ∞, based on the speech quality of the reconstituted speech from the summation, it has been found that only a finite number of coefficients are required in the representation. Since the signal $J_1([x_m/a]t)$ is bandlimited to $0 < |\omega| < x_m/a$ with large energy concentrated near $|\omega| = x_m/a$, the largest index m in the summation may be determined based on the highest frequency to be represented. To include no more than 3 kHz in the representation of a frame of speech, for example, the function $J_1([x_m/a]t)$ must be present in Eq. (1) with

$x_m = (2\pi a)3000 = 471.24$ for the range a = 25 ms.

Since the 150^{th} root of $J_1(t)$ is 472.02, the upper limit in the summation in Eq. (1) must be at least 150. To cover the entire Nyquist band of 4 kHz at the sampling rate of 8 kHz, $x_m = 628.32$ is needed; this is close to the 200^{th} root of 629.1.

Instead of computing all the 200 coefficients to represent each frame of speech to cover its frequency spectrum, a selected sample of coefficients can be used without loss of information. Although the coefficients may be selected based on their relative amplitudes or their frequency content [4,5], a set of coefficients that cover a bandpass frequency range was chosen with the indices given by m = 10, 20, 35, 50, 60, 70, 80, 90, 100, 115, 130, 145, 160, 175, and 190. At each of these indices, five successive coefficients were computed. These coefficients (15 x 5) represent a starting frequency of 205 Hz (m = 10) and go up to approximately 3905 Hz (m = 195). We may note that the frequencies at which the maximum spectral energies occur for the basis functions, $\{J_1([x_m/a]t)\}$, are somewhat analogous to the mel scale of frequency values chosen for the cepstral domain feature extraction. To reduce the feature vector size further, an energy measure in a narrow band of frequencies in the vicinity of each selected index was obtained. The magnitude of the basis signal amplitude at each of the five indices was added to get the energy measure as

$$e(k) = \sum_{m_k}^{m_k+4} |C\, m_k|, \quad \text{for } k = 1, 2, .. 15, \text{ and m from the set } \{10, 20, 35, 50, .., 190\} \qquad (3)$$

We note again that the feature element e(k) corresponds to the log energy output of the critical band filter in the cepstral domain representation. Thus the FB-based feature vector models the perceptual hearing at selected frequencies. By adding a fixed number of coefficients (five, for example, here), the bandwidth at each selected frequency is almost constant. Therefore, the comparison with the mel distributed cepstral coefficients based on the energy in a critical band of frequencies is complete if an increasing number of coefficients are considered in forming the feature vector e(k) in Eq. (3). The bandwidth corresponding to the five neighboring coefficients used in Eq. (3) is almost constant with the index m. At k = 1 (m = 10 to 14), for example, e(1) represents approximately the energy in the bandwidth of $f_1 = x_{10}/((2\pi a) = 204.93$ Hz, to $f_2 = x_{14}/((2\pi a) = 284.95$ Hz, or 80 Hz. At k = 10 (m = 115 to 119), e(10) represents the energy from $f_1 = 2305$ Hz to $f_2 = 2385$ Hz. Only the energy measure e(k) with 5 coefficients was used in constructing feature vectors for the Greenflag data base.

III. Identification Results

All the features considered in the previous section were tested using a vector quantizer-based classifier available from Entropic Systems. With the same classifier used on all the features and for all the data bases, the effectiveness of each feature may be compared in terms of computational complexity, robustness under noisy and stressful conditions and identification scores. We must note that the Greenflag data base used is noisy with bursts of engine noise and microphone clicks; also, many of the test utterances are very short in duration (a few hundred milliseconds) compared with the corresponding training utterances in the data base. With no endpoint detection, therefore, the burden of discriminating each speaker regardless of the utterance duration and the quality of speech is primarily on the features. The classifier

performance is not likely to adversely affect the identification score. Table I lists the scores for the Fourier transform-based features discussed.

Table I
Identification scores for the Greenflag data base using Fourier transform-based features

No	Feature	Score No. correct/Total	Score in percent
1	14 Cepstral Coeffts. [a]	99/173	57.2
2	19 Cepstral Coeffts [b]	112/173	64.7
3	50 from cepst. and lpc [c].	105/173	60.7
4	57 from cepst., lpc and diff. [d]	119/173	68.8
5	57 from cepst., diff., and ref. co [e].	121/173	69.9
6	20 Cepstral Coeffts [f]	139/173	80.4
7	20 Log energy values [f]	132/173	76.3

[a] The first 14 cepstral coefficients from the 19 given below.

[b] The cepstral coefficients were obtained with critical band filters centered at the frequencies of 102 Hz, 203 Hz, 305 Hz, 406 Hz, 508 Hz, 609 Hz, 711 Hz, 813 Hz, 914 Hz, 1016 Hz, 1148 Hz, 1320 Hz, 1516 Hz, 1734 Hz, 1992 Hz, 2289 Hz, 2633 Hz, 3023 Hz, and 3469 Hz.

[c] 10 parameters from each of lpc, reflection coefficients, cepstrum, three-frame differential cepstrum and five-frame frame differential cepstrum.

[d] 19 parameters from each of cepstrum, three-frame differential cepstrum and lpc

[e] 19 parameters from each of cepstrum, five-frame differential cepstrum and reflection coefficients

[f] 20 cepstral coefficients and the log energy evaluated using 2048-point DFT at critical band filters and bandwidths different from those of previous computations. The approximate center frequencies of the filters are: 116 Hz, 198 Hz, 300 Hz, 398 Hz, 500 Hz, 600 Hz, 700 Hz, 800 Hz, 898 Hz, 1000 Hz, 1148 Hz, 1316 Hz, 1508 Hz, 1738 Hz, 1996 Hz, 2288 Hz, 2628 Hz, 3020 Hz, 3460 Hz, and 3672 Hz.

As expected, because of the close modeling of the human auditory perception, the mel-based cepstral coefficients performed better than the lpc-based parameters. The differential cepstral coefficients, regardless of the frame difference, did not raise the score when combined with the cepstral coefficients to form feature vectors. Neither did the lpc representation help raise the score when used with other features. Using a higher frequency resolution (by increasing the number of points in the DFT computation) and using a slightly different set of points for the cepstrum computation, a 20-point cepstral representation increased the identification score significantly. In addition, the log energy at the cepstral points (outputs of the 20 critical band filters before the IDFT operation in Fig. 1) yielded an identification score comparable to that of the cepstral feature.

Based on the above results, we may conclude that the cepstral domain features are more suitable for speaker identification when the signals have significant amount of noise and the utterances have varying durations. The speed of identification may be increased at the cost of a slightly lower score by using the spectral energy as the feature element since this avoids the inverse DFT

operation. From the two sets of cepstral features, it is also clear that the choice of the cepstral indices to form feature vectors may depend on the data base. If the data base, unlike the Greenflag, had female voices, for example, the cepstra may need to include more high frequency components.

As with the cepstrum-based features, the identification results obtained using the FB transform-based features depended on the coefficient indices used. A set of five consecutive raw coefficients at seven (and, later, eight) points over the range of m = 1 to 200, for example, performed poorly in representing the Greenflag data base and yielded an identification score of below 30 %. This is analogous to using the spectral amplitudes at selected frequencies, which may not form a good representation for the speech or the speaker, depending on the choice of the frequencies. The energy measure given by Eq. (3), as with the cepstrum, resulted in reasonably high scores as shown in Table II.

Table II

Identification scores for the Greenflag data base using FB transform-based features

No	Feature	Score No. correct/Total	Score in percent
1	15 Energy values at selected indices	128	74
2	20 Energy values at selected indices	131/173	75.6
3.	20 Differential energy values at selected indices - frame diff.	81/173	46.8
4	19 Differential energy values at selected indices - index diff.	122/173	70.5

The slightly higher score for the 20 parameter feature vector could be attributed to the choice of indices more than to the increased vector size. As with the cepstral coefficients, this choice depends on the speaker set. If the utterance duration is small, as is the case with some of the test utterances in the Greenflag data base, the indices must be carefully selected to bring out the few available key features.

The last two rows in Table II show the results based on feature vectors that are similar to the delta and the differential cepstra - the item in the third row uses the two-frame difference in the energy measure, i.e., for the m^{th} frame, df(m,n) = e(m,n) - e(m+1,n), n = 1, 2, .. 20; the last item uses the difference in the energy measures between two adjacent groups of indices, i.e., d(m,n) = e(m, n+1) - e(m,n), n = 1, 2, .. 19. Clearly, the differential energy between adjacent groups of indices gives results comparable to those of the energy values. It would be informative to determine if the two sets of parameters, the energy values and the differential index energy values, correctly identify the same speakers. If, on the other hand, the ordering of the distortion

measures from the classifier using the two feature sets does not differ significantly, a fusion of the two sets may be used to get a better identification result.

The features discussed above were also tested on a second data base, the NATO data base, using the same classifier. The NATO data base consisted of transmissions of speech utterances from nine speakers (pilots). Although the number of speakers is small, the total number of utterances available was significantly higher (1117) than any other data base with transmissions from aircraft. Therefore, the result of any speaker identification process on this data base would be significant. The utterances were represented using each of the parameters discussed above and speaker identification tests were performed using the same VQ classifier. Although the duration of each transmission was long (typically about 2.5 s), most of it contained only background noise. Therefore, seven transmissions were concatenated together for each speaker to obtain a reasonable length of speech in each reference utterance. (When four transmissions were used as reference for each speaker, the identification test resulted in poor scores.) Since the transmissions were obtained at a sampling rate of 16000/s, initially a frame length of 512 samples for the cepstrum-based features and 400 samples for the FB transform-based features were chosen. These choices for the frame length were based upon the tests conducted on the Greenflag data base, where 200 samples/frame was used for the transmissions obtained at 8000/s. The following table (Table III) lists the identification scores obtained for the 1054 = 1117 - 7x9 test transmissions.

Table III
Identification scores for the NATO data base

No	Feature	Score No. correct/Total	Score in percent
1	20 Cepstral Coeffts.	928/1054	88.0
2	20 Log energy values	885/1054	84.0
3	20 Energy values using FBC at selected indices -- 400 samples/frame	582/1054	55.2
4	20 Differential energy values using FBC at selected indices - frame diff. -- 400 samples/frame	467/1054	46.7
5	20 Energy values using FBC at selected indices -- 200 samples/frame	683/1054	64.8
6	20 Differential energy values using FBC at selected indices - index diff. -- 200 samples/frame	651/1054	61.8

As the above table indicates, the cepstral coefficients and the log energy at selected frequencies with varying bandwidths clearly performed better than the FB transform-based features. A

reason for the low score using the FB transform-based features at 400 samples/frame is the low frequency range represented by the FB coefficients that formed the features. At the indices of the coefficients used for the representation, namely, {5, 15, 25, 35, .. 185, 195}, the frequencies covered are only from approximately 105 Hz to 3905 Hz. While this range of frequencies is sufficient to represent speech, the cepstral domain features cover a much wider range of 233 Hz (at the first selected cepstral index of 26 with 2048-point DFT and 512 points/frame) to 7344 Hz (at the last - 20^{th} - index of 940). With the high recognition score using this range of frequencies in the cepstral domain features, it is clear that other features must include at least the same range. Therefore, the frame length for use with the FB transform was changed to 200 samples. At this frame length, the range became 12.5 ms and the frequency coverage changed to 210 Hz to 7810 Hz. The last two rows show a significant increase in the identification score with the increased frequency coverage.

It is expected that with a judicious choice of the coefficients used for the representation (instead of the arbitrary choice at intervals of 10 coefficients), the scores would improve further. In addition, a better representation of hearing perception would result with overlapping coefficients in the formation of the energy measure given in Eq. (3). The increased number of coefficients would result in increased computational effort, however.

IV Analysis of Stressed Speech

The goal of this part of the project was to determine a set of parameters obtained from a speech utterance that would indicate the level of the speaker's stress. The NATO data base contained utterances from nine pilots with corresponding transcription files for each speaker. The transcription file for a speaker included information on the measured stress levels and the number of speech transmissions at each level. With the measured stress level varying only by a small percentage, however, it was decided to consider another data set that was part of the NATO data base for the initial analysis. This data set consists of mayday transmissions for two pilots and ground crew.

The transmissions for the pilot who lost his engine were extracted from the combined transmissions of the pilots and the ground crew. The following is the transcription of the extracted utterance by the pilot:

Hydraulic oil pressure light is still lit; I've lost my engine mayday mayday mayday; No gyro vectors let's go; Stopping right turn; Put the cable down put the cable down; God damn; Yeah I don't have an engine; Copy that guys, thanks for all your help I owe you; That's it newt it stopped.

The above transcription indicates that the stress level of the pilot may be elevated during the mayday transmission. However, we should note that in the absence of normal or baseline transmission for the same pilot, it may not be possible to compare the variation of any feature considered for analysis.

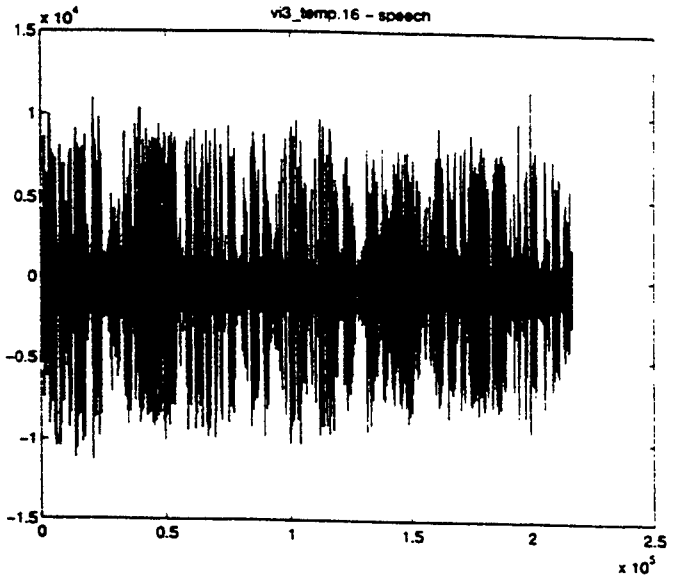
It is generally observed [6, 7, 8, 9] that in addition to the spectral tilt, other acoustical attributes of fear, anger, agitation and work stress are the fundamental frequency of voicing F0, and the first three formants F1, F2 and F3. Since the Bessel functions are better suited for voiced speech, initially the energy measure given by Eq. (3) was used to represent the extracted mayday speech. Fig. 2 shows the plots of the signal and the energy in the 20 sets of selected coefficients that cover the frequencies (at starting indices) of 210 Hz (5), 609 Hz (15), 1010 Hz (25), 1410 Hz (35), 1810 Hz (45), 2210 Hz (55), 2610 Hz (65), 3010 Hz (75), 3410 Hz (85), 3810 Hz (95), 4210 Hz (105), 4610 Hz (115), 5010 Hz (125), 5412 Hz (135), 5810 Hz (145), 6210 Hz (155), 6610 Hz (165), 7010 Hz (175), 7410 Hz (185) and 7810 Hz (195).

The energy plot shows variation of the spectral energy at low coefficient indices. Since the first 10 values, which show large variation, correspond to frequencies from 210 Hz to 3810 Hz, it is clear that energy plot (Fig. 2b) and the differential energy plot (Fig. 2c) both reflect the variation in F0 to F3. Without a reference utterance of the same speaker at low stress level, however, we cannot ascertain how much of the FB-based parameter variation is due to change in stress level and how much is due to linguistic content and speaker-related attributes. .

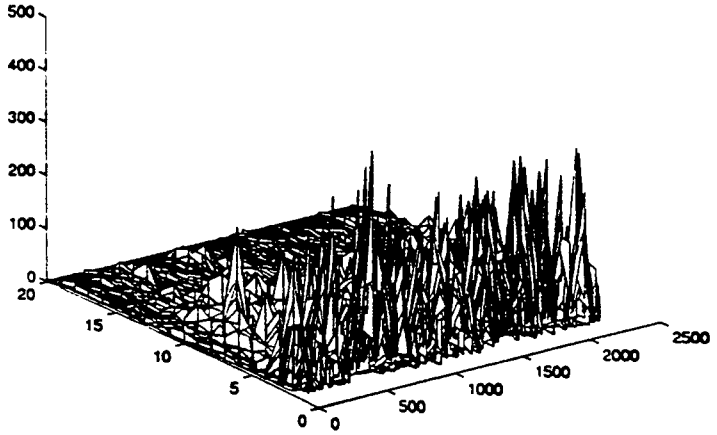
For the fundamental frequency, which is observed to vary from 50 Hz to 400 Hz [7, 9] the appropriate range of coefficient indices to consider are from 1 to 10. Fig. 3. shows a plot of the first 10 coefficients for the mayday utterance. Clearly there is a large variation in C_1 - C_5 near the 1000^{th} frame, and another significant variation in C_6 - C_9 in the vicinity of the 500^{th} frame. Although other frames also show deviations in their coefficients, particularly in C_7 - C_9, the lower indices, from 1 to 5, are most likely indicating the fundamental frequency variation for the male voice analyzed. Further processing is needed to understand how much of the variation is due to the mayday-induced stress.

The 20 cepstral coefficients at the indices that were used to form features for speaker identification did not show any discernible variation for the mayday speech. As seen in Fig. 4, this choice of the indices (corresponding to the frequencies 116 Hz, 198 Hz, 300 Hz, 398 Hz, 500 Hz, 600 Hz, 700 Hz, 800 Hz, 898 Hz, 1000 Hz, 1148 Hz, 1316 Hz, 1508 Hz, 1738 Hz, 1996 Hz, 2288 Hz, 2628 Hz, 3020 Hz, 3460 Hz, and 3672 Hz) may not be suitable to indicate the changes due to stress. In particular, since F0 variation is the most prominent one in stressed speech, the coefficients must reflect the range of 50 Hz to 400 Hz.

(a)



(b)



(c)

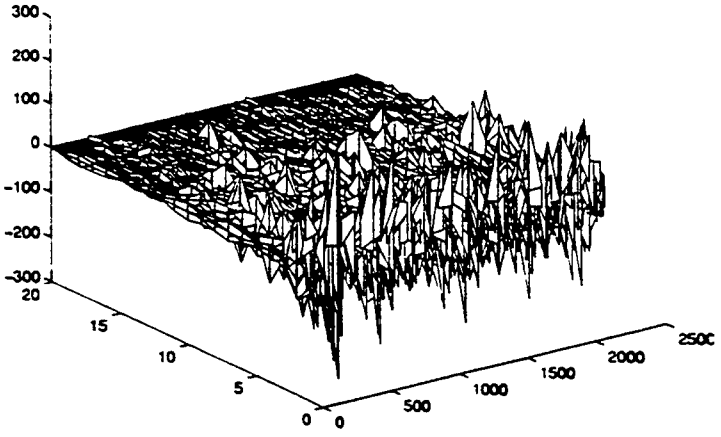
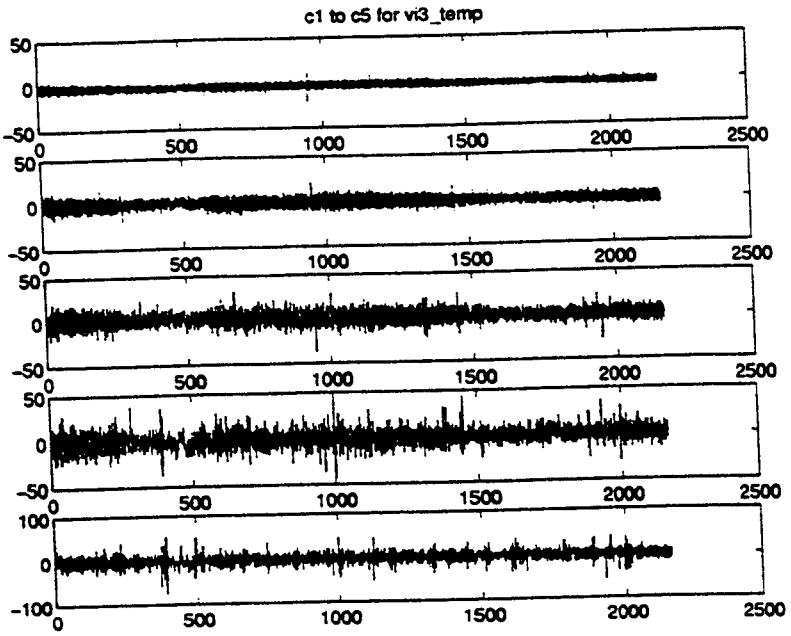


Fig. 2 (a) Mayday utterance (b) FB-based energy plot at selected indices (c) Energy difference between successive frames



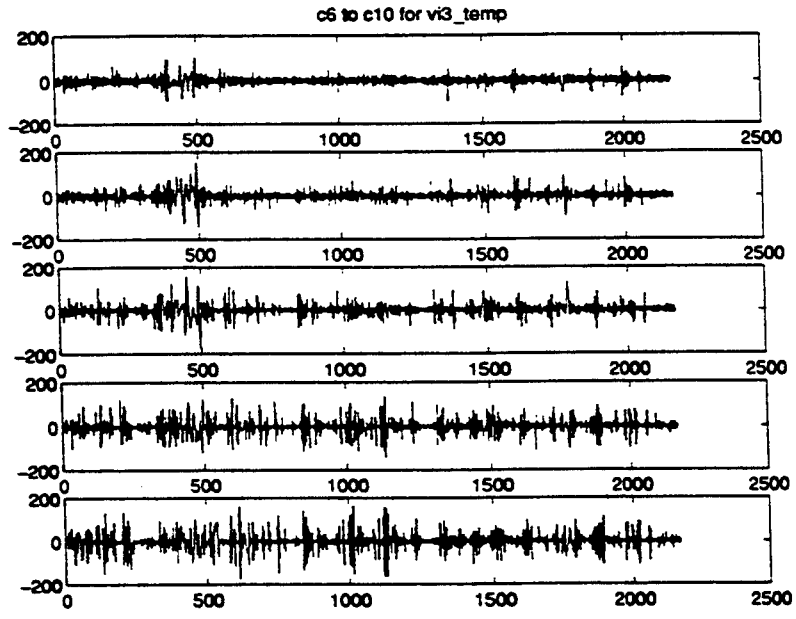
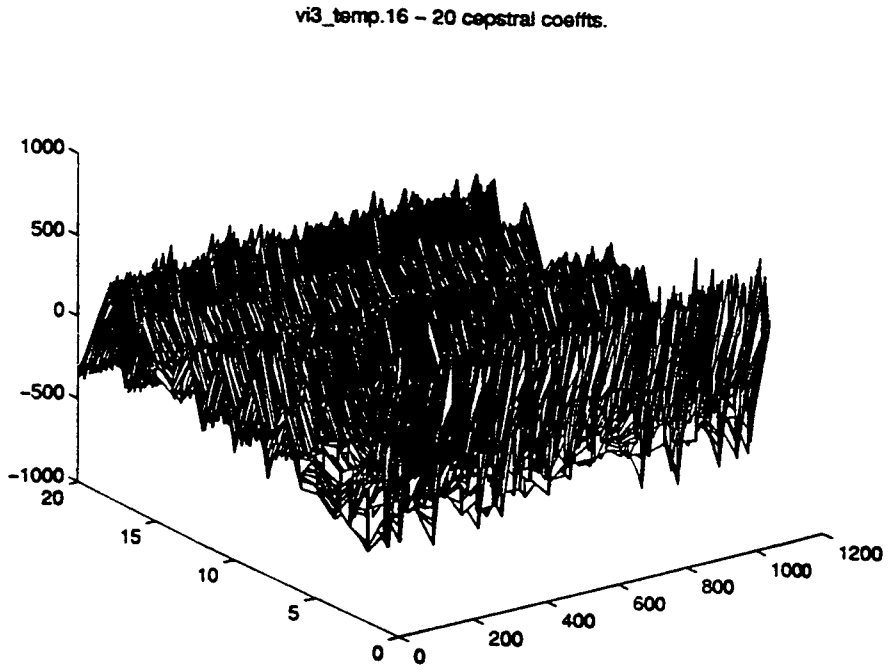


Fig. 3 Plot of the first 10 Fourier-Bessel coefficients C_1 - C_{10} in the representation of the mayday speech

(a)



(b)

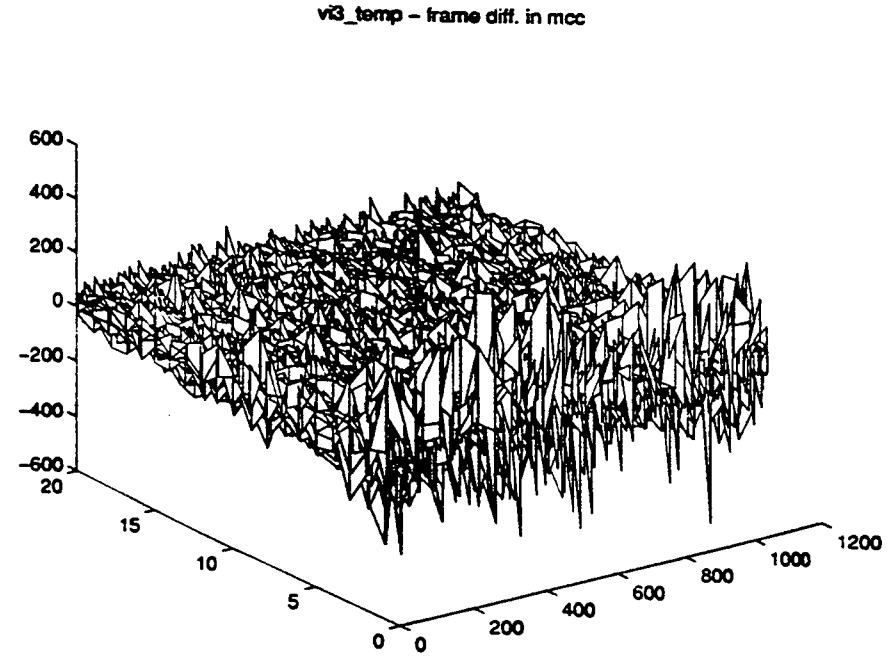


Fig. 4 (a) Cepstral coefficients at 20 selected indices and (b) their successive frame differences

Wavelet-based signal representation was also considered for the analysis of stressed speech. In the preliminary study the mayday utterance was filtered by a Morlet wavelet. With the bandpass response of the wavelet centered at 150 Hz, the filtered speech indicated signal components around the fundamental frequency. By analyzing the energy of the filtered speech, therefore, a measure of the variation in F0 may be obtained. In addition, with a sharp characteristic of the wavelet spectrum, the filtered speech at various time scales of the wavelet may be used to determine the trajectory of F0. The preliminary study of the filtered speech and the energy in each frame showed that the choice of the wavelet bandwidth is crucial to the analysis of the stress-related characteristics.

Other features, such as those resulting from the zero crossing rate and the zero crossing integral after Morlet wavelet filtering of the signal were computed for analysis. These features were not pursued further due to lack of time.

V. Further work

Fourier-Bessel representation offers an alternative to Fourier transform-based features for speech processing. For improved modeling of hearing perception and acoustic waveform, more coefficients in the representation may be needed. To compare with the cepstral features used in speaker identification, as stated in the previous section, the coefficient indices may need to overlap with the neighboring values; also, more coefficients are needed for feature extraction to increase the bandwidth at higher indices. Additionally, when the signal has large duration of background noise, it may help to detect the endpoints and use only the speech segment for representation. This will reduce the data size and the processing time.

For stressed speech analysis, it may be possible to relate the coefficient indices with the fundamental frequency and formants if the coefficient variations are studied with known stress levels.

The high identification scores achieved by the cepstral features for both the Greenflag and the NATO data bases indicate that cepstral processing may bring out the stress-related acoustic parameters better in the feature domain than the FB transform-based features. The choice of the cepstral indices for this purpose may be different from the indices used for identification. Variation of the cepstral coefficients and the log energy with frames must be studied for obtaining a profile of speaker-specific linguistic and non-linguistic characteristics. Additionally, the choice of window used with the frames is important in preserving the nonstationary spectral behavior and the model for hearing perception. From the recent experiments on using nontraditional triangular-type windows [10], a suitable window for obtaining the stress correlates may be the one with linear rise and exponential fall [11]. Since the ability of this window in phoneme recognition was verified, it may also provide a better model for stressed speech analysis. Effect of this window on cepstral coefficients and liftered cepstra must be analyzed.

With the recent developments on the use of orthogonal wavelets for speech recognition, pitch determination and speaker identification [12, 13], wavelet-based multiresolution analysis may provide a measure of the stress-related acoustic features. The choice of the wavelets and the level of resolution need to be determined.

VI. Conclusions

Speaker identification on large data bases of speech transmissions show that a small set of 15 to 20 cepstral coefficients yield the highest scores compared with linear predictive parameters and delta cepstrum. Slightly lower scores result from the use of an arbitrary set of coefficients in the construction of features using the Fourier-Bessel expansion. These scores demonstrate that the Bessel functions may be an alternative to the conventional sinusoidal basis functions for representing speech signals.

Preliminary study of a stressed speech utterance shows that features such as the Fourier-Bessel expansion and the cepstral coefficients at different frequencies may provide a measure of stress-related acoustic parameter variations.

References

[1] S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE T. ASSP, Vol 28, No. 4, Aug. 1980, pp. 357 - 366.

[2] J.R. Deller, J.G. Proakis and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, New York, 1993

[3] C.S. Chen, K. Gopalan and P. Mitra, "Speech Signal Analysis and Synthesis via Fourier-Bessel Representation," Proc. ICASSP, Tampa, FL, March 1985, pp. 497-500.

[4] K. Gopalan, "Speaker Identification using Bessel Function Expansion of Speech Signals," Final Report for Summer Faculty Research Program, Armstrong Laboratory, AFOSR, August 1993.

[5.] K. Gopalan and T. R. Anderson, "Speech Processing using Bessel Functions, Proc. Symp. Intelligent Systems in Communications and Power, Mayaguez, PR, Feb. 1994, pp. 255-259.

[6] J.H.L. Hansen, B.D. Wiomack and L.M. Arslan, "A Source Generator Based Production Model for Environmental Robustness in Speech Recognition," Proc. ICSLP, 1994, pp. 18.5-118.5-4

[7] C.E. Williams and K.N. Stevens, "Emotions and Speech: Some Acoustical Correlates," JASA, Vol. 52, No. 4 (Part 2), 1972, pp. 1239-1250.

[8] D.A. Cairns and J.H.L. Hansen, "Nonlinear Analysis and Classification of Speech under Stressed Conditions," JASA, 96 (6), Dec. 1994, pp. 3392-3400.

[9] D.R. Ladd, et al, "Evidence for the Independent Functions of Intonation Contour Type, Voice Quality, and F0 Range in Signaling Speaker Affect," JASA, 78(2), Aug. 1985, pp. 435-444.

[10] B. L. Losiewicz, "Windowing Comparison Project: The Effect of Window Shape and Size on Phoneme Identifiability," Rome Laboratory Report, Summer 1991.

[11] E.J. Cupples - Personal Communication.

[12] M. Bodruzzaman, et al, "Speaker Recognition using Neural Network and Adaptive Wavelet Transform," Proc. SPIE, Vol. 1961, Orlando, April 1993.

[13] S. Kadambe, et al, "Representation and Classification of Unvoiced Sounds using Adaptive Wavelets," Vol. 1961, Orlando, April 1993.

Mode-locked Laser Models and Simulations

J. W. Haus
Physics Dept.
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

Final report for:
Summer Faculty Research Program
Rome Laboratory, Rome, NY

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC
and
Rome Laboratory

October 1996

Mode-locked Laser Models and Simulations

J. W. Haus
Physics Dept.
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

Abstract

Models for mode-locked lasers were developed and computer simulations of the models were made. The lasers we examined operate in the 1.5 μm wavelength regime, which is of interest for C3I applications. I was especially interested in the characteristics of three lasers at the Photonics Center of Rome Laboratories: the Cr^{4+}:YAG laser, the harmonic mode-locked fiber laser and the passive, semiconductor mode-locked fiber laser. The Cr^{4+}:YAG laser model has been successfully modeled with results that help understand the experimental laser conditions. Modeling is continuing with the other two laser designs.

One papers was prepared for publication and a second is being planned, based on experiments with mode-locked fiber lasers. Conference papers were delivered at the Optical Society of America Conference and an Air Force workshop in Tucson, AR. An abstract was submitted to the Optical Fiber Conference, which will be held in February, 1997.

1 Introduction

To generate short, i.e. sub-picosecond pulses, modelocked laser designs are used[1]. In the past we have modeled modelocked fiber lasers based upon the fast saturable aborber-type action of a nonlinear optical loop mirror[2, 3]; these are called figure-eight lasers because of the shape of the path traced out by the signal. Our modeling reproduced effects that were previously observed in experiments. Our understanding of the operation of these lasers also directed us to improve the design by managing dispersion in the cavity. This resulted in a version of the laser that we dubbed the *balanced figure-eight laser*[3].

Solid state materials with wide fluorescence emission spectra have been exploited in a number of laser applications to generate optical femtosecond pulses[4]. To accomplish ultrashort pulse operation a number of mode-locking techniques using artificial saturable absorber have been developed; Kerr lens mode-locking relies on changes in the pulse's transverse profile to modulate the cavity gain [4] and a saturable absorber mirror incorporates a saturable absorber on one cavity mirror to modulate the cavity loss [5, 6].

Erbium-doped fiber lasers (EDFL) are also very useful for producing short pulses. They have several cavity designs and operation regimes, eg. continuous wave, mode-locking and Q-switching regimes; they are being explored as a source of high repetition rate, energetic, ultrashort pulses, that meets the needs of future system applications.

Ultrashort pulses are generated by mode-locking techniques. Active mode-locking, using, for example, a Mach-Zehnder modulator in the cavity, has been valuable as a source of regularly spaced, high repetition-rate pulses [7]. Passive mode-locking, using a fast saturable absorber, has produced pulses of sub-picosecond duration. The pulses in each case are soliton-like, i.e. hyperbolic-secant shaped; solitons have proven to be robust against the presence of losses and amplification in fiber transmission systems; i.e. they don't alter their shape in the presence of small perturbations. The topic of soliton transmission in optical fibers has rapidly evolved from a pure research topic to an emerging technology through a series of important technological breakthroughs (overcoming challenging obstacles) in long-distance, high bit-rate communication systems. In addition, soliton interactions have been proposed for logic and routing devices, which perform important information processing tasks [8].

Recently self-starting, passive mode-locking of a Cr^{4+}:YAG laser using a saturable absorber mirror (SAM) was reported[9, 10]. The SAM is designed from a quarter-wave stack of GaAs/AlAs layers with a double quantum well structure grown at the interface. This is an important element in the cavity design. Its absorption changes as the laser is tuned, which in turn affects the cavity mode-locking action. The successful operation of the laser depends upon the specific features of this element. The prism pair in the cavity provides negative dispersion for solitonic pulse shaping in the cavity. The laser operating wavelengths are in the range from 1490 nm to 1540 nm, which lies in the transmission region for fiber optic communications. The output power varies between 40 and 80 mW, which is intense enough to launch pulses in fibers with sufficient energy to study multi-soliton propagation effects with ultra-short pulses.

Using a simple model we are able to accurately predict the mode-locking behavior of the Cr^{4+}:YAG laser. The cavity design is reduced to a small number of parameters that are independently measured or determined.

2 Cr^{4+}:YAG Model

The average soliton approach is viable to describe the properties of a Cr^{4+}:YAG laser. The equation in this regime is[11]

$$\frac{\partial a}{\partial z} = i(-D\frac{\partial^2 a}{\partial t^2} + \delta|a|^2 a) + D_3\frac{\partial^3 a}{\partial t^3} + (g-l)a + \frac{g}{{\Omega_f}^2}\frac{\partial^2 a}{\partial t^2} + \gamma_3|a|^2 a - \gamma_5|a|^4 a. \quad (1)$$

This equation is a form of the complex Ginzburg-Landau equation studied in several fields of physics. The amplitude a represents the complex electric field envelope. D is the cavity dispersion, which is related to the second-order dispersion β_2 by $D = \frac{\beta_2 L_D}{2}$; L_D is the length of the medium that contributes to the dispersion; the third order dispersion coefficient is defined by $D_3 = \beta_3 L_D/6$. The Kerr nonlinearity, n_2, responsible for self-phase modulation, appears in the parameter $\delta = \frac{2\pi}{\lambda} n_2 P L_x/A_{eff}$; P is the peak power; A_{eff} is the effective area of the beam; λ is the wavelength; and L_x is the crystal length. The scaled value for the nonlinear parameter is $\delta = 0.8$. The cavity loss and gain is collected in the parameters l and g, respectively. The gain bandwidth is parameterized by Ω_f, whose value is $\Omega_f = 2.9 \cdot 10^{14} Hz$; and the saturable absorber is described by two parameters γ_3 and γ_5. The formation of bright solitons requires the second-order dispersion be negative; this is fulfilled by the dispersion compensating prisms. The first two parameters in parentheses are the soliton-shaping mechanisms of dispersion and nonlinearity.

The dominant pulse shaping mechanism in the cavity are due to dispersion and self-phase modulation terms. The form of the pulse amplitude is a hyperbolic-secant function

$$a = \eta \mathrm{sech}(t/\tau) e^{iz/2L_D}. \quad (2)$$

$L_D = \tau^2/2|D|$ is called the dispersion length; for positive δ the dispersion is negative for bright solitons. The gain and saturable absorber terms determine the pulse energy, $W = 2\eta^2\tau$, while these and higher-order dispersion terms perturb the pulse shape from the soliton solution. Important parameters are the average pulse width and the pulse energy, which are closely related by the dominant shaping terms:

$$\tau = 4|D|/\delta W. \quad (3)$$

The pulse shape found by our simulations is close to a hyperbolic-secant form with deviations appearing in the wings. The energy is calculated from the area under the pulse. The balance between the energy in the linear and nonlinear gain-loss contributions to the energy is given by

$$l - g = \frac{2}{3}\gamma_3\eta^2 - \frac{8}{15}\gamma_5\eta^4 - \frac{1}{3}\frac{g}{{\Omega_f}^2\tau^2}. \quad (4)$$

The last two terms, which are higher-order absorption saturation and gain dispersion terms, resp., make important corrections to this result and our results are consistent with this expression. This equation can be solved for the pulse width and the energy to determine consistency of our soliton shaping hypothesis represented by Eq. (3); this too is very closely followed and this consistency check is evaluated in the results. The deviations can be attributed to the third-order dispersion, which contributed to the continuum radiation.

The numerical computations were done using a split-step algorithm, which is described elsewhere and employed for our earlier studies of pulse propagation [12, 13, 14]. The linear coefficients were chosen as $l = 0.02$ and $g < l$ is adjusted to keep the steady-state pulse energy constant. Eq. (4) imposes restrictions on the maximum value of l; this is discussed later. In the experiments, the output coupler must be kept small to achieve mode-locking. Group delay dispersion and the third-order dispersion in the cavity are computed applying the techniques. The values for the cavity were calculated for the design in Ref. [10] and are tabulated in Table 1. The dispersion calculations for the prism pair can be found in literature[15].

Table 1

wavelength (nm)	second-order dispersion fs^2	third-order dispersion fs^3
1470	-1679	12061
1480	-1796	12372
1490	-1915	12691
1500	-2037	13019
1510	-2160	13356
1520	-2286	13703
1530	-2413	14058
1540	-2543	14424
1550	-2675	14799

The quintic-order terms, γ_5, in Eq. (1) was required for stable operation. Without it the peak energy was unstable. This is also observed in Eq. (4), where the cubic term, γ_3, is balanced against the quintic term. The gain dispersion in the cavity is also important in the energy balance.

3 Results

The experimental results for the pulse width and spectral width are reproduced from Ref.[10] and shown in Fig. 1. The trend in the data over the range of frequencies is mostly attributable to the variations in the dispersion, although there are also variations in the output energy. However, the curvature at longer wavelengths toward longer pulse widths exceeds expectations based on variations of the dispersion alone. This can be accounted for by the wavelength dependence of the absorption. The semiconductor material operates near the band edge and the absorption changes as the wavelength is tuned. The change can be rather large and is modeled in Figure 2. The roll off of the absorption strongly affects the ability of the laser to achieve pulse operation.

Our simulation results did not exhibit a rapid rise in the pulse width, as observed in the experiment, Fig. 1. The results were qualitatively correct, but the pulse width change was not as dramatic in the simulations. The pulse width curve did not significantly change when third-order dispersion, also provided in the Table, is added to the simulation. The cavity gain was

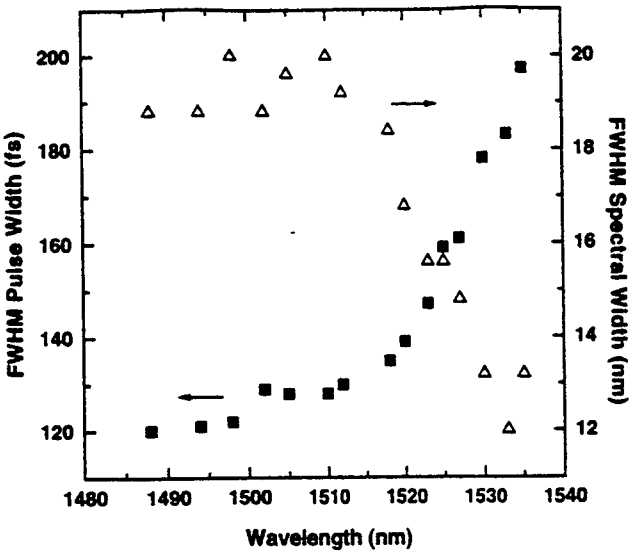


Figure 1: Experimental curves for the pulse width and spectral width versus wavelength. All values are full width at half maximum. After Ref. [10].

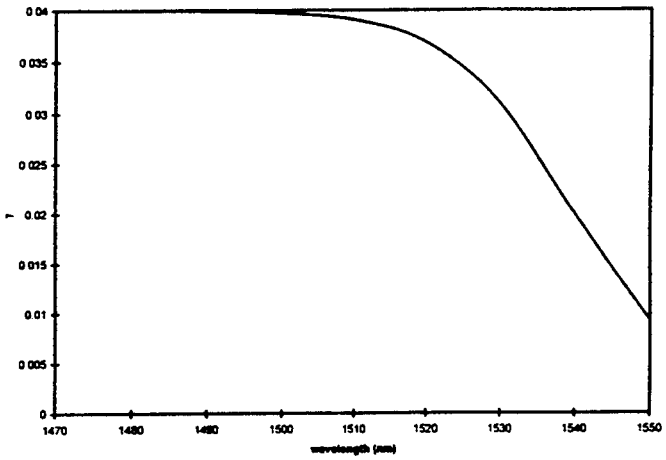


Figure 2: The variation of the absorption saturation curve with wavelength using Eq. (4) and the parameter values quoted in the subsequent text.

adjusted at each frequency to keep the steady-state energy of the pulse constant. For constant pulse energy the pulse width is approximated from Eq. (3) and the results are indistinguishable from the simulation points. This is clearly shows the dominant soliton shaping mechanism for our model. The pulse width is estimated from the simulation is also in agreement with Eq. (4); deviations are within 6%. We believe that the differences can be attributed to the third-order dispersion, which increases due to the cavity gain.

Simulations were refined to study the cause for the sharp upward trend of the pulse width as the laser was tuned to longer wavelengths, see Fig. 1. Initially the sensitivity of the results to parameter variations was examined; the energy was held constant in these simulations by adjusting the gain. The loss was varied between 0.1 and 0.2; the gain bandwidth was changed between 2.9×10^{14} Hz and 1.93×10^{14} Hz; and the saturable absorber parameter given values of 0.02 and 0.04. The change in the gain needed to hold the energy constant did not change the trend of the pulse width versus wavelength; the pulse width was well described by Eq. (3).

The excess change in the pulse width in the tuning curve of Fig. 1 is attributable to the change in the saturable absorber efficiency as a function of wavelength. This causes a variation of the output energy, which is consistent with the sharp upward turn of the pulse width at longer wavelengths. The absorption varies near the band edge becoming less effective for wavelengths below the band edge. The absorption was measured as a function of wavelength. The nonlinear coefficients are fit to a function of the following form, we chose $\gamma_5 = \gamma_3$ in units scaled to the field saturation intensity

$$\gamma_3 = \frac{0.04}{1 + \exp(E_g - E)/\sigma}. \tag{5}$$

The parameters chosen for the function are $\sigma = k_B T_w$ and $E_g = hc/\lambda_g$, where h, c and k_B are Planck's constant, the speed of light in vacuum and the Boltzmann constant, resp. The wavelength $\lambda_g = 1500nm$ and the effective temperature width is $T_w = 50K$. These are chosen to correspond to the experimental parameters for the saturable absorber mirror. The the saturation coefficient versus wavelength is shown in Fig. 2; the value above the gap is taken as 0.04.

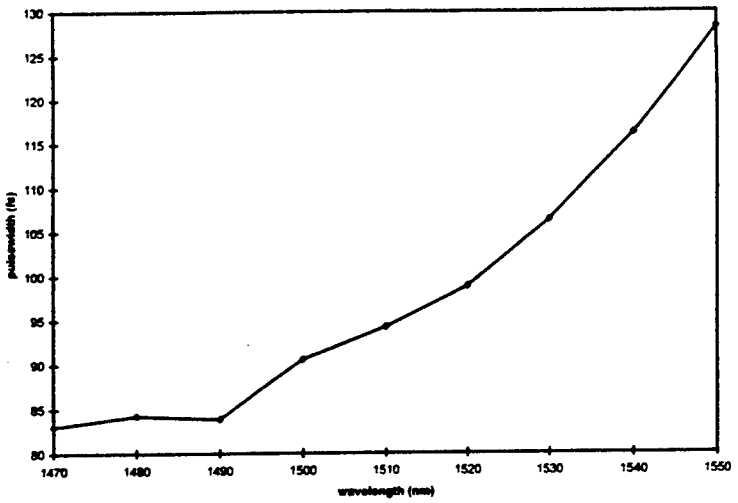


Figure 3: The pulse width versus frequency from the simulations the solid line excludes the variation in absorption saturation; and the dashed line includes the variation in saturation.

When the saturation variation is taken into account, the results are modified. The dashed curve of the pulse width versus frequency in Fig. 3 shows the effect that this has on the pulse width. The decrease in the saturable absorber parameters becomes severe at the long wavelengths, which results in a slow convergence to the steady-state solution. We were unable to find mode-locking for wavelengths longer than 1550 nm, just as in the experiment. The output energy cannot be held constant in these simulation by merely adjusting the gain parameter g. The corresponding change of the output energy with wavelength is displayed in Fig. 4.

Figs. 5 and 6 are plots of the temporal and spectral pulse profiles versus frequency. the central portions of the temporal pulse profile are well approximated by a hyperbolic secant shape with an excess background of radiation in the wings. The spectrum has oscillations in the central region due to perturbations affecting the soliton propagation, but its wing remains exponential over four orders of magnitude.

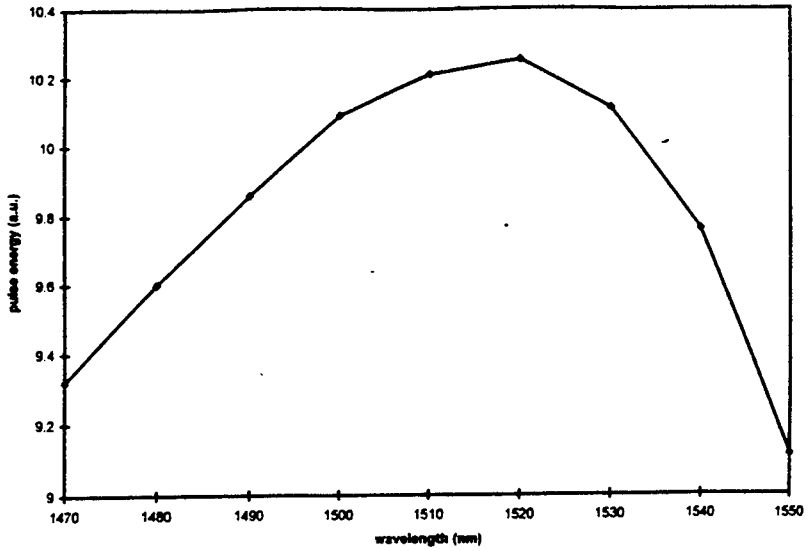


Figure 4: The energy of the steady-state pulses versus wavelength using the change of the absorption coefficient. The peak at 1520 nm occurs at the knee of the absorption curve in Fig. 3.

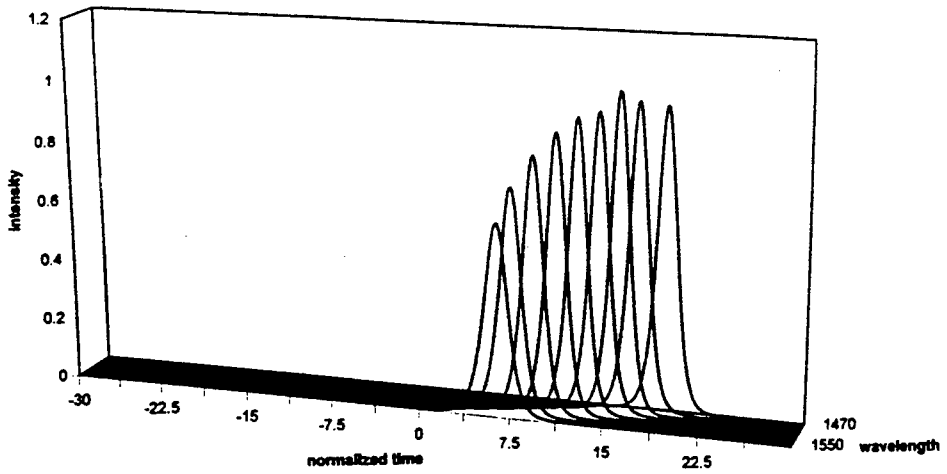


Figure 5: Pulse profile versus time and tuning wavelength for the steady-state pulses.

4 Summary

The average soliton methods previously developed for solid–state lasers has been applied to the saturable-absorbing mirror geometry of the Cr:YAG laser. The computational simulations reproduce the trend observed in experiments[10] in the pulse width versus tuning curve, when the variation of the saturation parameters with wavelength is taken into account. Third-order dispersion is not significant for the cavity design of Ref. [10].

Our analysis of Eq. (4) indicates that tolerable cavity losses are limited by the effectiveness of the saturable absorber. A gross upper limit is $l_{max} = 4\gamma_3\Omega_f^2|D|/\delta$, which varies from about 0.2 to 0.05 for our numerical simulations (depending on the value of γ_3). Further limits on the loss in Eq. (4) are imposed by demanding the pulse widths are real and positive.

The experimental situation is more restrictive than our simulations since the bandwidth is compressed by the tuning slits in the cavity dispersion section, which filter the radiation, and by wavelength dependent changes in the mirror reflectivity which becomes smaller at longer wavelengths. This, together with the reduced saturable absorber coefficients at long wavelengths, are

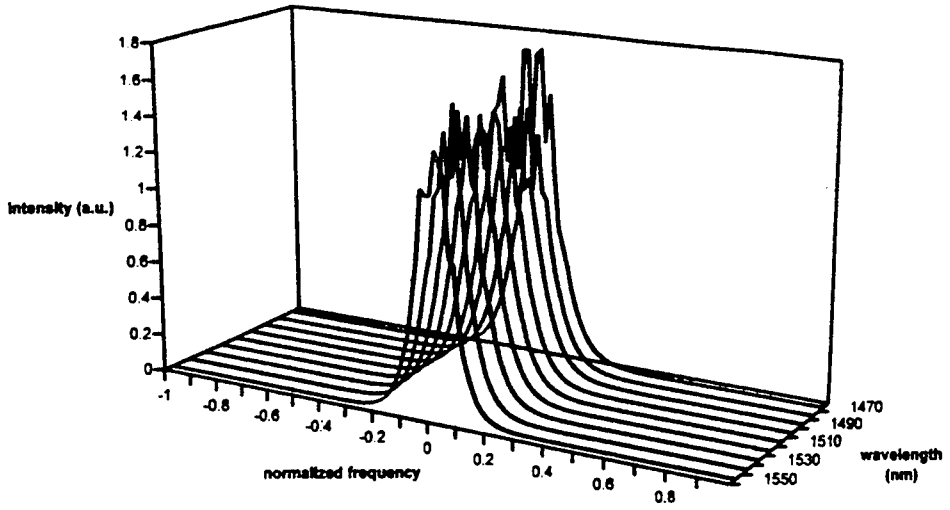


Figure 6: Pulse spectrum versus frequency and tuning wavelength for the steady-state pulses.

responsible for the eventual inability of the laser to mode-lock beyond 1540 nm in the present design. This was observed both in the experiments and in the simulations. Our results indicate that the Cr:YAG laser can produce shorter pulses at the longer wavelengths by redesigning the saturable absorber mirror to shift the band gap to longer wavelengths and by having smaller prism dispersion in the cavity and by mirrors designed for higher reflectivity and lower loss at the longer wavelengths. For the design in Ref.[10], the third-order dispersion made no significant contribution and by reducing the group velocity dispersion, the pulse width could be further reduced.

5 Research Activities

Building on past experience with mode-locked fiber lasers we have been able to successfully model the operation of the Cr^{4+}:YAG solid state laser. We find that the saturable absorber mirror is a critical element in the laser design. In the Photonics Center laser the efficiency of the saturable absorber rolls off at long wavelengths due to the position of the semiconductor band edge.

Work is continuing to understand the harmonically mode-locked fiber laser and the passively mode-locked fiber laser. Both are being studied by my student, Walter Kaechele, who was also a summer student at the Laboratory. He has conducted a series of experiments on both lasers, whose interpretation requires modeling and simulation studies. Furthermore, noise characterization and the effect of input laser noise on the synchronization of both lasers needs to be determined. This will be important for future applications to local area networks using a time-division multiplexing architecture.

The preliminary results of our research on the Cr^{4+}:YAG were presented at an Air Force sponsored workshop in October, 1996. A manuscript is being prepared for submission to Optics Communications; it contains the details of the modeling and a comparison between our simulations and the experimental results. Research is continuing on the relevant fiber lasers that are now operating in the laboratory by Walter Kaechele.

Acknowledgments

The support of the Photonics Center and especially, of James Theimer, Andrew Pirich and James Cusack is gratefully acknowledged. Without their help this research could not have been com-

pleted.

References

[1] M. J. F. DIGONNET ed., *Rare Earth Doped Fiber Lasers and Amplifiers*, Marcel Dekker, New York (1993).

[2] J. Theimer and J. W. Haus, "Figure-eight laser stable operating regimes," J. Mod. Optics, in press (1996).

[3] J. Theimer and J. W. Haus, "Dispersion Balanced Figure-eight Laser," Opt. Commun., in press (1996).

[4] T. Brabec, M. J. Kelly and F. Krausz, "Passive modelocking in solid state lasers," see I. N. Duling III ed., pp.57-92.

[5] D. Kopf, G. Zhang, R. Fluck, M. Mosser and U. Keller, Opt. Lett. **21**, 486 (1996).

[6] S. Tsuda, W. H. Knox, E. A. de Souza, W. Y. Jan and J. E. Cunningham, Opt. Lett. **20**, 1406 (1995).

[7] G. T. Harvey and L. F. Mollenauer, "Harmonically modelocked fiber ring laser with an internal Fabry-Perot stabilizer for soliton transmission," Opt. Lett. **18**, 107-109 (1993).

[8] M. ISLAM, *Ultrafast Fiber Switching Devices and Systems*, Cambridge University Press, Cambridge, (1992).

[9] B. C. Collings, J. B. Stark, S. Tsuda, W. H. Knox, J. E. Cunningham, W. Y. Jan, R. Pathak and K. Bergman, Opt. Lett. **21**, 1171-73 (1996).

[10] M. J. Hayduk, S. T. Johns, M. F. Krol, C. R. Pollock and R. P. Leavitt, preprint, (1996).

[11] H. A. Haus, "Short pulse generation", in I. N. Duling, III ed., Compact Sources of Ultrashort Pulses, (Cambridge Univ. Press, Cambridge, 1995), pp 1-56. H. A. Haus, J. G. Fujimoto and E. P. Ippen, "Analytic theory of additive pulse and Kerr lens mode locking," IEEE J. Quant. Electron. **28**, 2080-96 (1992).

[12] G. P. Agrawal, "Nonlinear Fiber Optics", (Academic Press, NY, 1989).

[13] J. W. Haus and J. Theimer, "Polarization Distortion in Birefringent Optical Fibers," Photonics Technology Letters**7**, 296 (1995).

[14] J. Theimer and J. W. Haus, "Figure-eight Fiber Laser Stable Operating Regimes," J. Mod. Optics , to appear (1996).

[15] R. L. Fork, O. E. Martinez and J. P. Gordon, "Negative dispersion using pairs of prisms," Opt. Lett. **9**, 150-3 (1984).

MULTICHANNEL AUTOREGRESSIVE MODELING AND SPECTRAL ESTIMATION METHODS FOR AIRBORNE RADAR ENVIRONMENT

James P. LeBlanc
Assistant Professor
Klipsch School of Electrical and Computer Engineering

New Mexico State University
Thomas and Brown Hall
Las Cruces, NM 88003-8001

Final Report for:
Summer Faculty Research Program
Rome Development Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Development Labs

October, 1996

MULTICHANNEL AUTOREGRESSIVE MODELING AND SPECTRAL ESTIMATION METHODS FOR AIRBORNE RADAR ENVIRONMENT

James P. LeBlanc
Assistant Professor
Klipsch School of Electrical and Computer Engineering
New Mexico State University

Abstract

The ability of multichannel AR models to properly model the received signal in an airborne radar environment was investigated. A physical model was used for generation of noise, clutter, and signal returns. Two different methods of the multichannel AR parameter identification were used, solution of the standard Yule-Walker and Overdetermined Normal Equations Methods. Results show that AR models of modest order well match the 2D power spectrum (computed by the 2D Fourier transform of the received data matrix) of the radar returns. The implications of acceptable modeling performance might indicate successful operation of innovations based detection algorithms (IBDA) in similar radar scenarios.

1 Introduction

A well-known detection technique in radar systems is the *estimator-correlator* receiver which first estimates the signal components and then correlates these with the received data. This type of processing, while effective, is computationally intensive. Furthermore, when operating within unknown signal statistics the problem is exacerbated. An alternative approach is to use parametric based methods, such as an *innovations based detection algorithm.* In this method, the received signal is converted to an innovations process. Detection is then based on the parameterization of the achieved model (such as by a log likelihood ratio test). In a true implementation the conversion to an innovations process (and the associated parameter estimation) may be done in an iterative fashion using computationally efficient adaptive filters.

The performance of such methods will depend upon a variety of things. First, there are the convergence issues associated with any adaptive system. Next, the particular innovations model chosen will also impact detection performance. This will be related to the operating environment of the radar system, and is influenced by:

- radar system parameters (frequencies, number of array elements, etc.)
- platform motion
- clutter environment
- presence and characteristics of interference
- presence and characteristics of targets

Fundamentally, one may ask "How well does a chosen parametric (innovations) model reflect the radar environment?" This is the main question addressed within this report. We seek to identify how well a chosen parametric model performs.

The actual "performance" metric of any surveillance radar system should be "how well the method detects targets". Realize that such a choice of metric would require analysis (and simulation) of the processing through the detection phase for a variety of radar scenarios. The amount of effort entailed would be well beyond the scope of the present research. Therefore, in this report we use a more tractable metric of ascertaining the model performance. Here we consider the model's ability to match the 2-dimensional (bearing and doppler space) power spectrum for given scenario consisting of simulated target, clutter, interference and noise signals.

We find that parametric models of modest order may model well the investigated radar sce-

narios. Additionally, a parameterization which yields acceptable performance may be estimated from a relatively small number of samples. Such qualities offer experimental evidence for the existence of computationally efficient parametric based detection algorithms within the airborne radar environment.

2 Radar Parameters and Scenario

The operational environment of radar systems vary widely. Our interest herein is airborne surveillance radar, the operational environment model and radar parameters are chosen to reflect such a case. Specifically, we use the radar system parameters as listed in Appendix A, Table 1, within the clutter environment modeled as listed in Table 2. Further details of these variables and models may be found in [5]. For all simulation runs the radar system parameters were held constant as were all signal geometries (i.e. target locations, motion, and target strengths, bearing of broadband interference and clutter model). Only the absence or presence of clutter, targets, and jammers differed in the various test runs.

Figure 1 depicts an idealized 2-dimensional spectrum of the scenario under consideration. The two vertical stripes correspond to broadband jammers at spatial frequencies of $f_s = \{0.211,\ -0.321\}$. The three circles represent three targets at $f_t = \{0.350,\ -0.046,\ 0.333\}$ $f_s = \{-0.250,\ 0.354,\ 0.0\}$. The shaded ellipse represents the *clutter ridge* present in our model.

Although the methods used here may be easily modified to investigate other radar platforms, it is *not* clear that the same results would necessarily be attained. That is, our conclusion of successful low-order multichannel AR modeling is considered specific to the system addressed.

3 Mathematical Formulation

We begin with a mathematical description of the system under investigation. Although, there are various conventions in use for multidimensional systems, we choose to adopt notation as in [7].

The received radar signal from one transmit pulse consists of a N complex samples from each of the J array elements for each range cell. Thus, for a range cell, the received signal may be

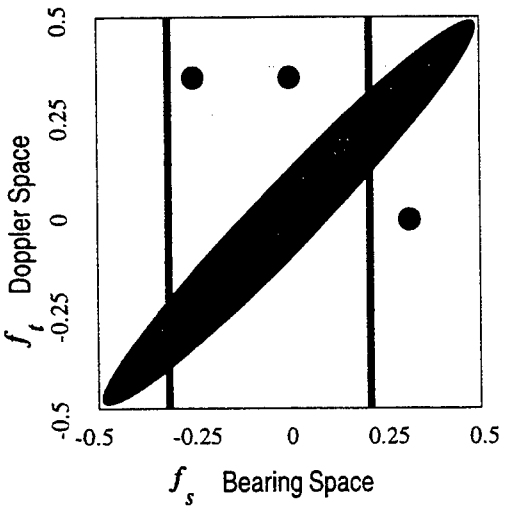


Figure 1: Radar Scenario

represented as a sequence of N, $J \times 1$ column vectors,

$$x(n) = \begin{bmatrix} x_1(n) \\ x_2(n) \\ \vdots \\ x_J(n) \end{bmatrix} \qquad 1 < n < N \tag{1}$$

where $x_i(n)$ represents the complex sample from sensor element i at time n. We assume that $x(n)$ is zero mean. In general, the received signal may be written as,

$$x(n) = s(n) + j(n) + c(n) + w(n) \tag{2}$$

suggesting that $x(n)$ contains:

$s(n)$	a signal component (target reflection return)
$j(n)$	broadband interference (jammers)
$c(n)$	non-white environmental backscatter (clutter)
$w(n)$	sensor and electronics white noise (noise)

The covariance matrix $R_{xx}(k)$ of a signal x is the $J \times J$ matrix,

$$R_{xx}(k) = \begin{bmatrix} r_{1,1}(k) & r_{1,2}(k) & \dots & r_{1,J}(k) \\ r_{2,1}(k) & r_{2,2}(k) & \dots & r_{2,J}(k) \\ \vdots & \vdots & \ddots & \vdots \\ r_{J,1}(k) & r_{J,2}(k) & \dots & r_{J,J}(k) \end{bmatrix} \tag{3}$$

where each element $r_{i,j}(k)$ is the correlation between channel i and j at lag k,

$$r_{i,j}(k) = \mathrm{E}\{x_i(n)^* x_j(n+k)\} \tag{4}$$

where * represent the complex conjugate. In general, $R(k)$ is not hermitian (except for $k = 0$) but does have the property that

$$R^H(k) = R(-k) \tag{5}$$

where ^{H} denotes the transpose and conjugate of a matrix.

When R_{ss}, R_{jj}, R_{cc}, and R_{ww} are known, a detection test based on comparing the estimate of R_{xx} (derived from the received signal — call this $\widehat{R}_{xx}$), under the hypotheses,

$$\begin{aligned} \mathcal{H}_0 &: x(n) = j(n) + c(n) + w(n) \\ \mathcal{H}_1 &: x(n) = s(n) + j(n) + c(n) + w(n) \end{aligned} \tag{6}$$

may be used. However, in a typical environment, one does not know a priori any of the correlation matrices and they also must be estimated from the received data. Such estimations and hypothesis testing requires considerable computing power. An airborne system places constraints on the available size/power/weight of computing machinery. Furthermore, a surveillance system, with its many "look directions" may require an order of magnitude more computations than a single beam radar system. For this reason, other detection algorithms need to be considered. One attractive method is the innovations based detection algorithm (IBDA).

4 Innovations Model

In IBDA methods, the received signal is converted to an innovations process (multichannel linear prediction process) as shown in Figure 2.

The filtering of $x(n)$ produces an output $\varepsilon(n)$, which has been whitened temporally as,

$$\varepsilon(n) = x(n) + \sum_{i=1}^{p} A_i x(n-k) \tag{7}$$

where each A_i is a $J \times J$ matrix. As mentioned previously the matrices A_i could be adapted in a computationally efficient method using adaptive filter to produce the "whitest" $\varepsilon(n)$. The idea

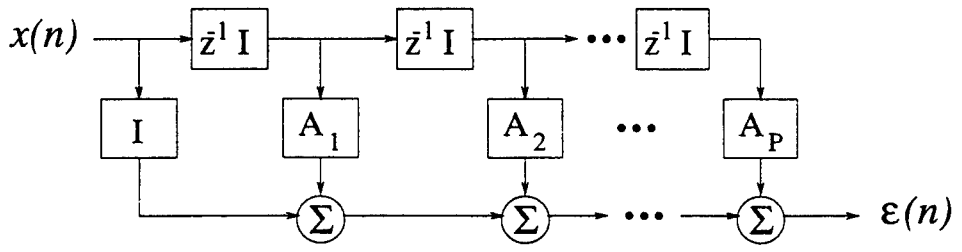


Figure 2: Innovations Process

behind IBDA is that the prediction filter coefficients A_i would then contain information about the received signal (clutter, jamming, signal), and hence a detection test could be based upon the achieved parameterization of the A_is. For a more complete description of the innovations approach see [3].

Necessarily, for the proper functioning of such algorithms is the assumption that different A_i parameterizations would be achieved under the target absent ($\mathcal{H}_0$ hypothesis), and the target present ($\mathcal{H}_1$ hypothesis). That is, for successful implementation one must have that signal $x(n)$ may be appropriately whitened by the filtering in 7. This is related to how well the process $x(n)$ may be *synthesized* by an order p autogressive (AR) process driven by (temporally) white (but spatial correlated) noise.

5 Multichannel AR Synthesis Model

The opposite of the innovations analysis filter (Figure 2) is the AR synthesis process in Figure 3. Let the input driving process, $u(n)$ be a temporally white $J \times 1$ vector process having correlation matrix given by,

$$R_{uu}(k) \;=\; \Sigma\delta(k). \tag{8}$$

where Σ is a $J \times J$ matrix of the cross-correlation between channels. The output of the synthesis filter is $x(n)$ is given by,

$$x(n) \;=\; -\sum_{i=1}^{p} A_i x(n-i) + u(n) \tag{9}$$

where each of the A_i are $J \times J$ matrices.

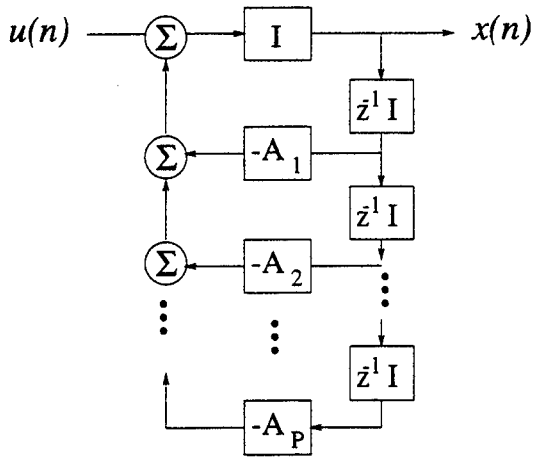


Figure 3: Multichannel AR Synthesis Process

By use of an innovations based detection algorithm, we may suspect there exists *some* appropriate synthesis model corresponding to our received signal $x(n)$.

What model order p is appropriate? How different are the the synthesis models under $\mathcal{H}_0$ and $\mathcal{H}_1$? It is these questions our investigation attempts to quantify. As mentioned earlier, full blown detection processing (a truly valid measure of performance) can not be addressed at this time. So, a simpler measure of model feasibility is pursued, the 2-D power spectrum.

6 Determining AR Model Parameters

We seek to establish the validity of AR models for IBDA processing for a radar scenario. This directs us to use a physically based model for the clutter, target and noise signals. The model chosen is describe fully in [5].

In actual implementation of IBDA, efficient adaptive filters would be used to determine the innovation filter parameterization (and hence the related synthesis filter). However, as here we desire to focus on model performance issues, we deliberately need to avoid issues relating to the adaptive filter's convergence.

6.1 *Square* Normal Equations

The model parameters (A_is and Σ) may be solved using the Yule-Walker equations. Let

$$\mathbb{R} = \begin{bmatrix} R_{xx}(0) & R_{xx}(-1) & \dots & R_{xx}(-(p-1)) \\ R_{xx}(1) & R_{xx}(0) & \dots & R_{xx}(-(p-2)) \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}(p-1) & R_{xx}(p-2) & \dots & R_{xx}(0) \end{bmatrix}.$$

Then we may write,

$$\begin{bmatrix} A_1^t \\ A_2^t \\ \vdots \\ A_p^t \end{bmatrix} = -\mathbb{R}^{-1} \begin{bmatrix} R_{xx}(1) \\ R_{xx}(2) \\ \vdots \\ R_{xx}(p) \end{bmatrix} \tag{10}$$

and $\Sigma = R_{xx}(0) + \sum_{i=1}^{p} R_{xx}(-i)A_i^t$.

However, lacking a closed form expression for the needed correlation matrices R_{xx}, we'll substitute estimated values $\widehat{R}_{xx}$ obtained from the received $x(n)$. To insure positive definiteness of $\widehat{\mathbb{R}_{xx}}$, typically the *biased* estimates of $r_{i,j}$ are used.

Solving for the AR parameterization in this way will be called the *square* method, since the block matrix of correlations ($\mathbb{R}$) is square, $Jp \times Jp$. This method is merely the multichannel version of a standard AR parameter determination strategy used in the scalar process case.

6.2 Overdetermined Normal Equations

Another method that has been proposed in the scalar process case is known as the *overdetermined normal equations* (ODNE) method [1]. It is extended here to the multichannel case. The basic idea is to make use of more of the estimated correlation lags than is allowed in the *square* normal equation method. That is, the *square* method inherently only uses correlation lags from 0 to p. In contrast, the ODNE method allows use of *more* estimated correlation lags by forming a non-square version of $\mathbb{R}$. This non-square version may consist of more than p estimated correlation lags and, as an estimate with more sample support, it may be possible to have a better solution. Here, we

form $\mathbb{R}$ as,

$$\mathbb{R} = \begin{bmatrix} R_{xx}(0) & R_{xx}(-1) & \dots & R_{xx}(-(p-1)) \\ R_{xx}(1) & R_{xx}(0) & \dots & R_{xx}(-(p-2)) \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}(p) & R_{xx}(p-1) & \dots & R_{xx}(1) \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}(t) & R_{xx}(t-1) & \dots & R_{xx}(t-(p-1)) \end{bmatrix}$$

where $t > p$, forming an overdetermined version of the system of equations of (10). Solution for the A_i may be found by use of the pseudoinverse of $\widehat{\mathbb{R}}$. In [1] it is mentioned that it may be useful in the ODNE method to use *unbiased* estimates of $r_{i,j}$. Both the square and ODNE solutions are investigated in this report.

7 Multichannel AR 2D Spectra

Given the driving noise correlation matrix Σ and model parameterization matrices A_i, the two dimensional spectrum of x may be computed as,

$$S_{xx}(f_t, f_s) = a^H \mathcal{A}^{-1}(f_t) e^{-j2\pi M f_s} \tag{11}$$

where, $a = \Sigma^{-1}[\sigma \quad 0 \quad 0 \quad \dots \quad 0]^t$, $\mathcal{A}(f_t) = I + \sum_{i=1}^{p} A_i^t e^{-j2\pi f_t i}$, and $M = [0 \quad 1 \quad \dots \quad J-1]^t$ from [4], and [8].

For comparison purposes, the resulting AR estimated 2D spectrum will be compared to the spectral estimate produce by taking the 2D, zero padded FFT of the matrix $\mathbb{X} = [x(1) \quad x(2) \quad \dots \quad x(N)]$. Also note that in such a classical form of spectral estimation, the data is often windowed by a matrix $\mathbb{W}$ before taking the FFT. In our experiment we display the results using both the windowed (using a Chebychev window of 60dB sidelobe) and unwindowed data.

When comparing the windowed FFT spectral estimate with the AR spectral estimate, it was found that solving (10) (or the ODNE version) using the windowed data to form $\widehat{\mathbb{R}}$ produced very similar spectra.

8 Results

The plots on the following pages display the results. Each scenario has an unwindowed and windowed data version. The power spectral densities for each method and scenario where average over 10 coherent processing intervals and plotted as an image where white corresponds to highest energy density, and black with the lowest energy density. Each Figure and consists of

- clutter and white noise only (Figures 4 & 7)
- clutter, targets, and white noise (Figures 5 & 8)
- clutter, targets, broadband jammers and white noise (Figures 6 & 9)

For each scenario,

- The top 9 plots show the power spectra resulting from the ODNE method using the *unbiased* correlation estimates.
- The next 3 plots show the power spectra resulting from the square method using *biased* correlation estimates.
- The bottom figure shows the periodogram (2D FFT of data).

The data plots are group columnwise by order of AR model. The first column uses $p = 2$, the middle column $p = 4$, and last column $p = 6$. Within the ODNE block various values of t (maximum lag used to form non-square $\mathbb{R}$) as denoted on the plots. The only difference between the second row (ODNE with $p = t$) and the third row (square) is the type of correlation estimate used (unbiased vs. biased).

9 Conclusion

The results of this research indicate that multichannel AR models of relatively modest order are capable of properly modeling the received radar signals in a simulated airborne radar environment. Here, the term "proper modeling" refers to matching the expected 2D power spectrum. Two methods of determining the AR parameters were investigated, the multichannel Yule-Walker equations, and a multichannel extension of the overdetermined normal equations. For most of the cases, the Yule-Walker equations using *unbiased* correlation estimates appeared to have the best performance. This is opposite of the single channel case where the ODNE method is often better. Perhaps a refinement is needed in the ODNE method in the multichannel case. More importantly, the proper

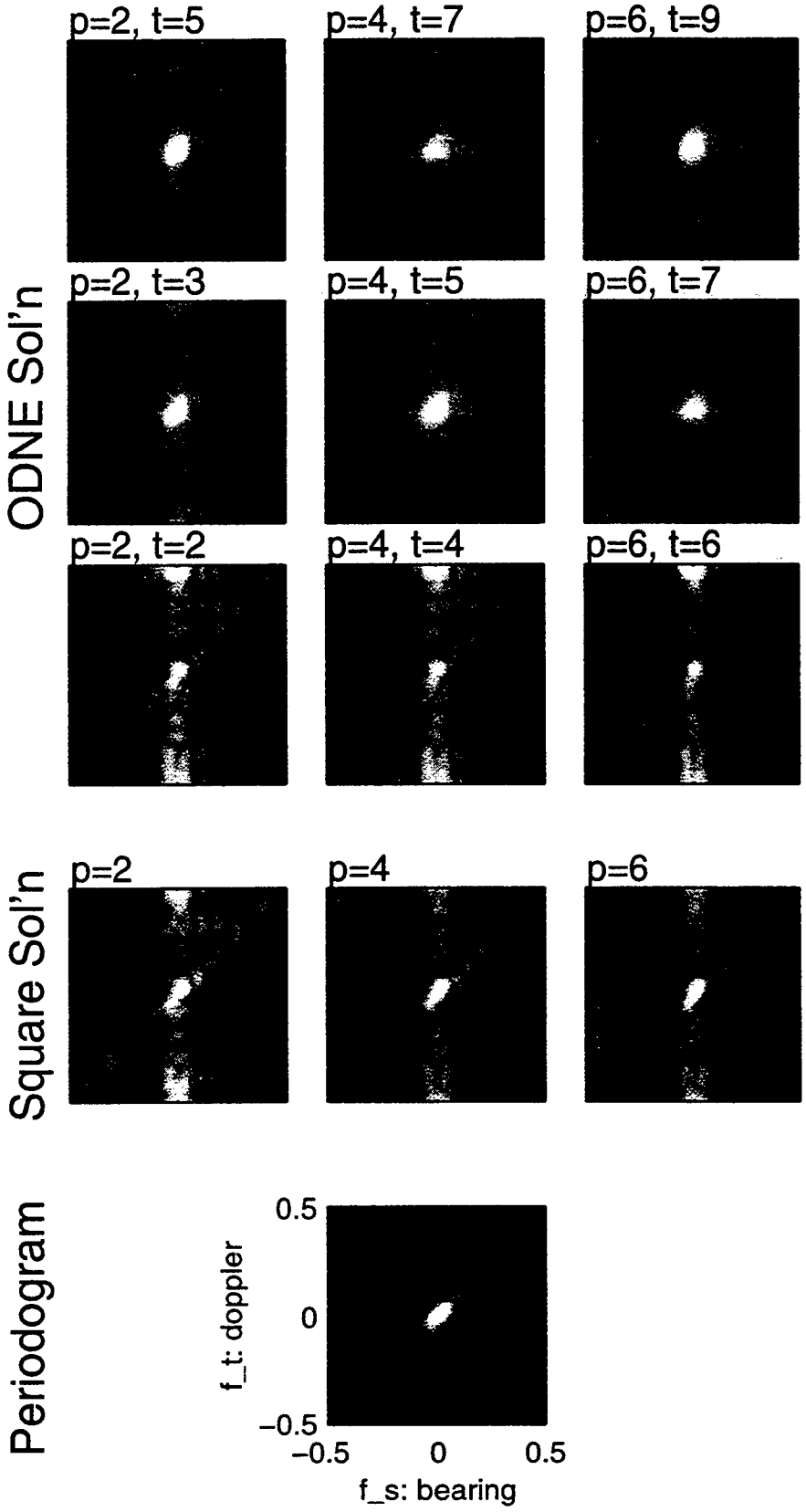


Figure 4: Non-Windowed Data: Clutter, White Noise

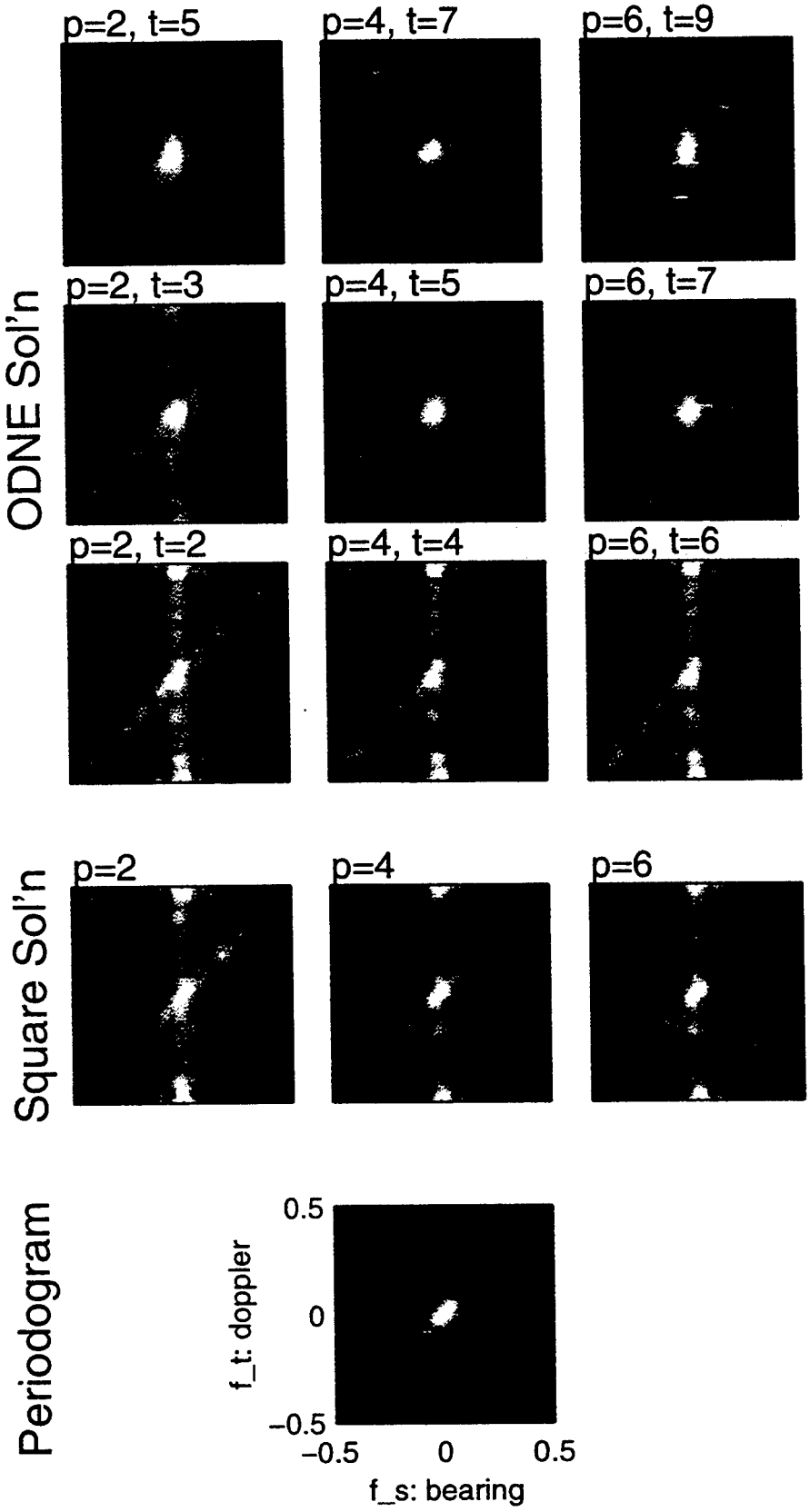


Figure 5: Non-Windowed Data: Clutter, Targets, & White Noise

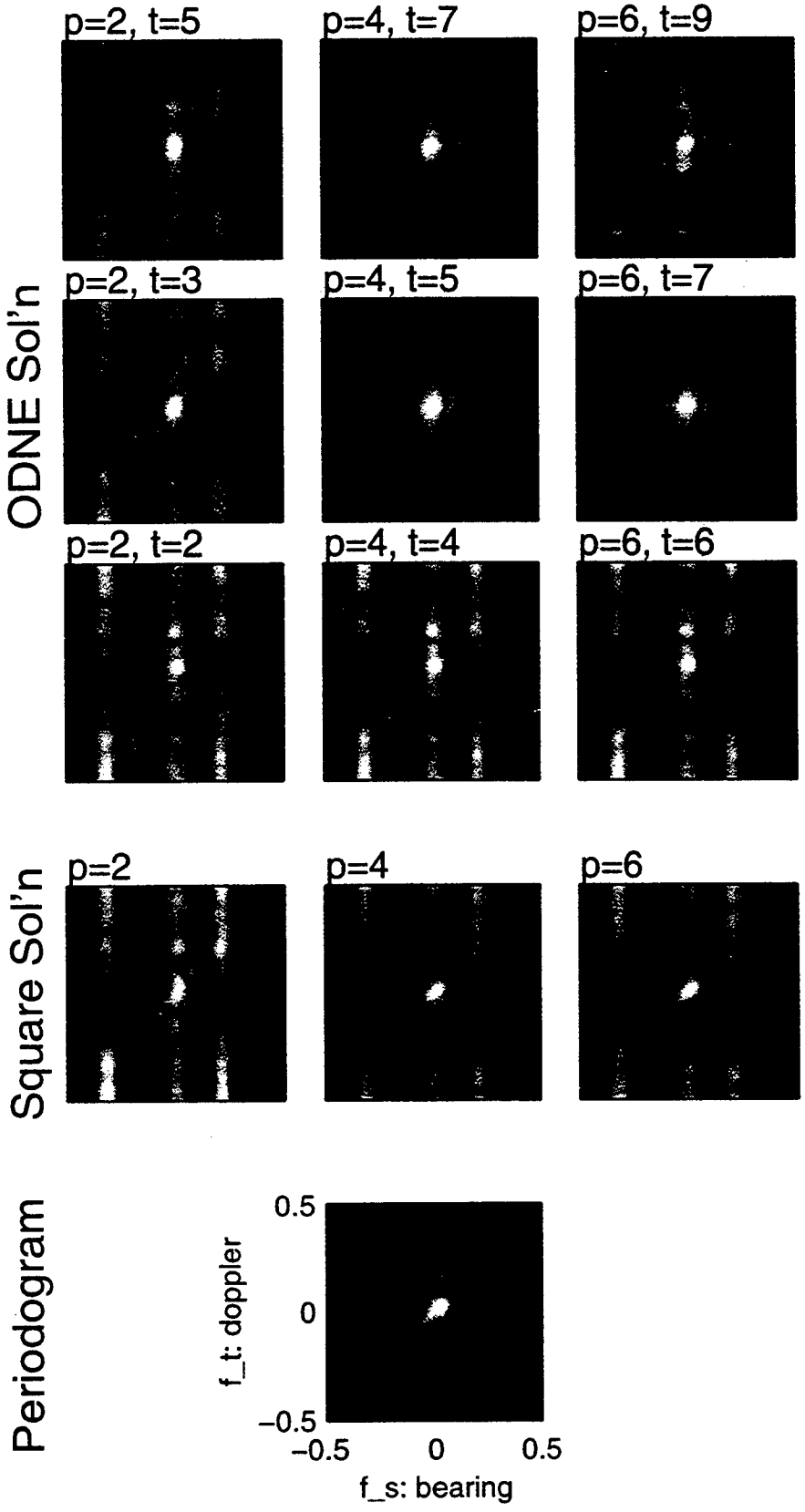


Figure 6: Non-Windowed Data: Clutter, Targets, Jammers, & White Noise

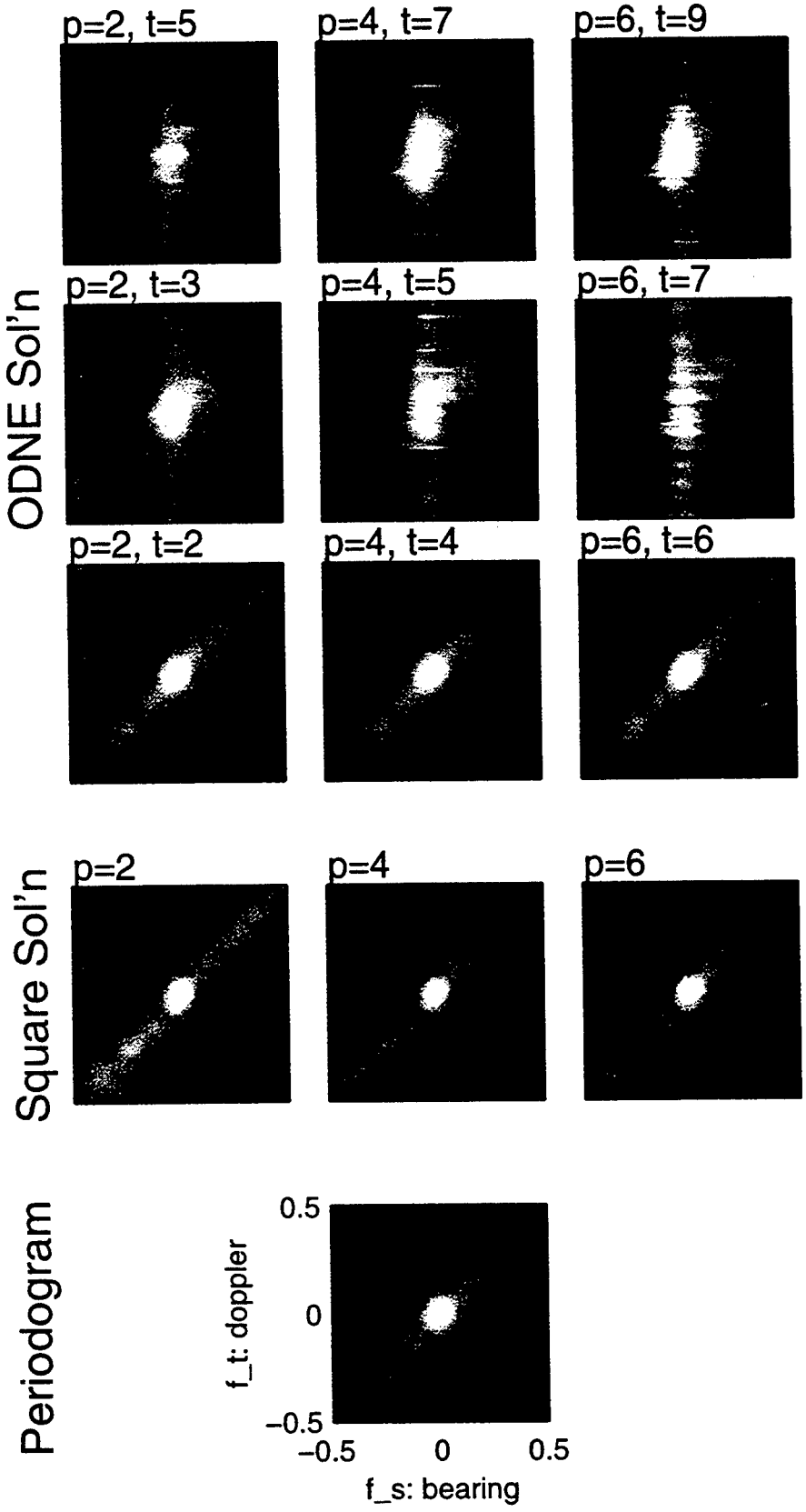


Figure 7: Windowed Data: Clutter, White Noise

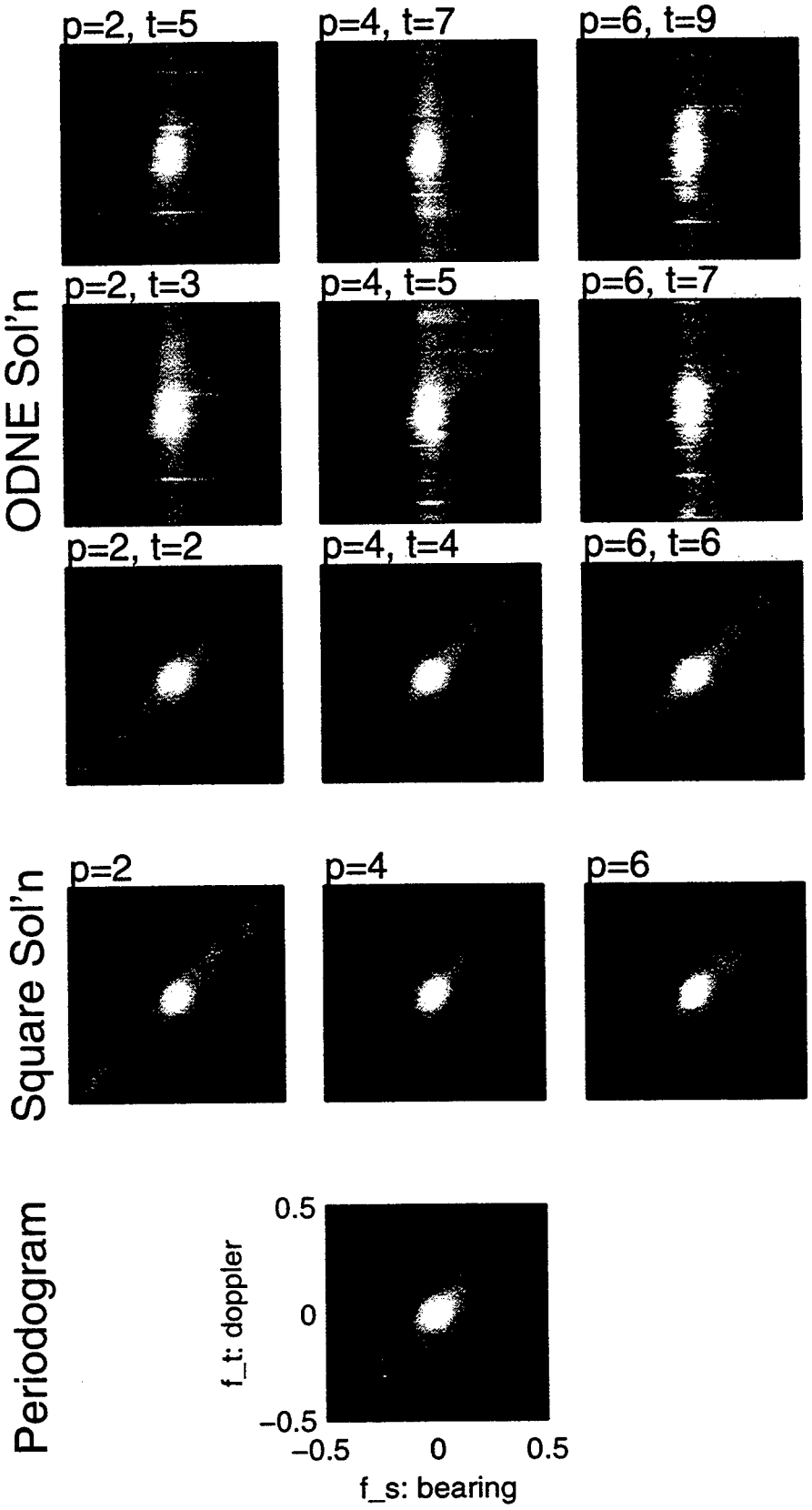


Figure 8: Windowed Data: Clutter, Targets, & White Noise

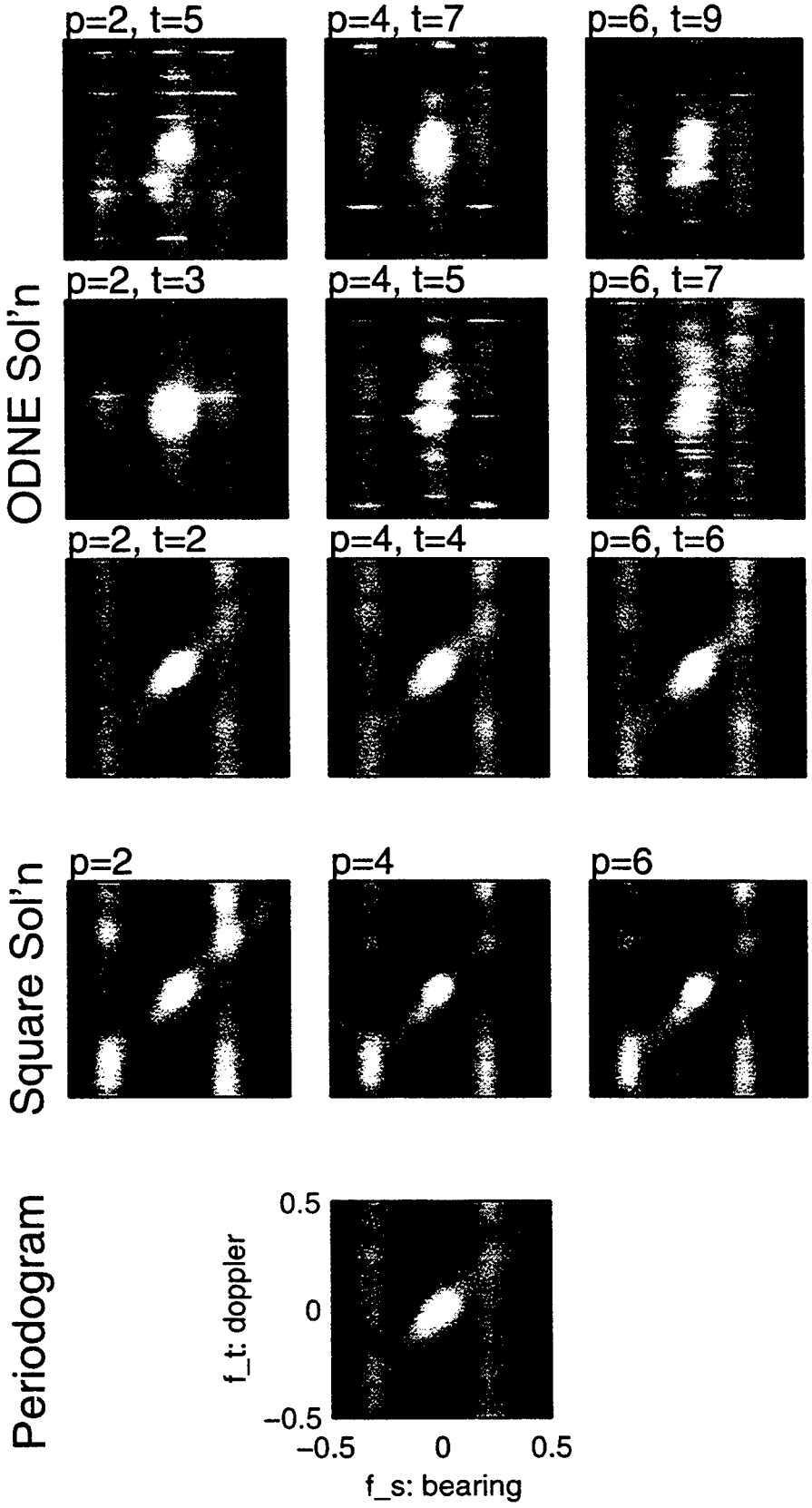


Figure 9: Windowed Data: Clutter, Targets, Jammers, & White Noise

modeling of radar returns by AR processes of relatively small order provides experimental support investigation of the use of multichannel IBDAs.

Appendix A — Radar System & Scenario Parameters

$J = 18$	Number of azimuth channels (azimuth axis array elements)
$N = 18$	Number of pulses in one coherent processing interval (CPI)
$J_e = 4$	Number of elevation axis elements beamformed into one azimuth channel
$\phi_0 = 0$	Array main beam azimuth angle, measured from boresight (deg)
$P_t = 200$	Peak transmitted power (kW)
$T_u = 200$	Pulse (uncompressed) duration (micro-sec)
$f_{PRF} = 300$	Pulse repetition frequency (Hz)
$f_C = 450$	Transmit frequency (MHz)
$f_B = 4$	Receiver bandwidth (MHz)
$G_o = 22$	Transmit pattern gain (dB)
$G_e = 4$	Receive element gain (dB)
$G_b = -30$	Receive element backlobe pattern attenuation (dB)
$F_n = 3$	Noise figure (dB)
$L_s = 4$	System losses (dB)
$txdcatt = 30$	Transmit pattern Dolph-Chebyshev weights sidelobe attenuation level (dB)
$patopt$ =UNIFORM	Array pattern option indicator

Table 1: Radar Array System Parameters

$H_p = 9$	Platform altitude (km)
$V_p = 50$	Platform velocity (m/s)
$r_c = 130$	Range to desired ground clutter ring (km)
$\gamma = 0$	Aircraft platform crab angle (deg)

Table 2: Surveillance Scenario Parameters

References

[1] J. A. Cadzow, "Spectral Estimation: An Overdetermined Rational Model Equation Approach," *Proc. of the IEEE*, Vol. 70, No. 9, Sept. 1982.

[2] J. H. Michels, "Multichannel Detection of Partially Correlated Signals in Clutter," *RL-TR-92-332*, Rome Laboratory, Air Force Materiel Command, Griffiss Air Force Base, New York, Dec. 1992.

TARGET PARAMETERS (SSC MODEL)	
a Damping Coef	{0.99, 0.99, 0.99}
f_s Spatial Frequency	{−0.25, 0.3536, 0.0}
f_t Temporal Frequency	{0.3502, −0.0465, 0.3336}
ϕ_t Azimuth	{−30, 45, 0} degrees
V_t Radial Velocity	{60, −40, 33.33} m/s
Signal Strengths	3dB
INTERFERENCE PARAMETERS (SSC MODEL)	
$G_i = 2$	Number of broadband interference sources (jammers)
$\phi_i = 25, -40$	Broadband random processes azimuth angles (deg)
$\theta_i = 0, 0$	Broadband random processes elevation angles (deg)
$var_i = 3310, 3000$	Broadband random processes powers
GROUND CLUTTER PARAMETERS (MIT/LL MODEL)	
$N_c = 361$	Number of ground clutter patches in the clutter ring
ARRAY NOISE PARAMETERS (SSC MODEL)	
$varn = 1$	Noise power in each channel

Table 3: Target, Clutter, Interference Parameters

[3] J. H. Michels, "Multichannel Detection using the Discrete-Time Model-Based Innovations Approach," *RL-TR-91-269*, Rome Laboratory, Air Force Systems Command, Griffiss Air Force Base, New York, Aug. 1991.

[4] C. W. Therrien, "Relations Between 2-D and Multichannel Linear Prediction," *IEEE Trans. on Acoust. Speech and Signal Proc.*, Vol. ASSP-29, No. 3, June, 1981.

[5] J. R. Roman, D. W. Davis, "Multichannel System Identification using Output Data Techniques Final Report No. SSC-TR-96-02," *Scientific Studies Corp., Palm Beach Gardens, FL*, March, 1996.

[6] S. Haykin, A. Steinhardt, *Adaptive Radar Detection and Estimation*, New York: John Wiley & Sons, Inc., 1992.

[7] S. Kay, *Modern Spectral Estimation*, Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[8] J. Roman, personal communication with J. Roman through J. Michels, August, 1991.

A REVIEW OF MICROWAVE TERRAIN CLUTTER MEASUREMENTS AT BISTATIC ANGLES

David J. McLaughlin
Associate Professor
Department of Electrical and Computer Engineering

Northeastern University
309 Dana Research
Boston MA 02115

Final Report for:
Summer Faculty Research Program
Rome Laboratory Hanscom

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory Hanscom

September 1996

A REVIEW OF MICROWAVE TERRAIN CLUTTER MEASUREMENTS AT BISTATIC ANGLES

David J. McLaughlin
Associate Professor
Department of Electrical and Computer Engineering
Northeastern University

Abstract

This paper reviews bistatic radar scattering measurements of terrain (clutter) surfaces that have been reported in the open literature. Brief descriptions of bistatic clutter measurement programs, conducted during the past three decades, are given. Normalized radar cross section (NRCS) values are tabulated and parameterized with respect to scattering geometry, radar frequency, polarization, and terrain type, and recommendations for future measurement programs, needed to extend the existing database, are given. This report summarizes the author's research performed over a 30 day period during the summer of 1996 as a participant in the AFOSR summer faculty research program at Rome Laboratory, Hanscom AFB, MA.

A REVIEW OF MICROWAVE TERRAIN CLUTTER MEASUREMENTS AT BISTATIC ANGLES

David J. McLaughlin

I. INTRODUCTION

Calibrated electromagnetic scattering measurements of terrain surfaces at bistatic angles are needed for estimating interference levels caused by: clutter in bistatic radars for target detection [1] and remote sensing [2]; terrain-scattered interference in monostatic radars subject to standoff jammer multipath [3,4]; and multipath signals in wireless communication systems. While numerous monostatic clutter measurements have been collected over the past several decades, relatively few bistatic terrain clutter measurements have been reported in the literature. Weiner [5] reviewed the bistatic clutter database collected prior to 1980, and Papa and Lennon [6] established the status of both monostatic and bistatic clutter phenomenology at L- and S-bands to 1981. Bistatic radar systems are currently being considered in a number of applications, and a number of clutter measurement programs have been conducted during the past decade to extend the existing database.

This paper reviews the bistatic clutter data that have been reported in the open literature to date. Part II summarizes the ten different experimental programs that have resulted in the current database of open-literature bistatic normalized radar cross section (NRCS) values for terrain clutter. This database was established using the bibliographies of [5] and [6] and by conducting CD-ROM library searches of National Technical Information Service (NTIS), Scientific Citation Index (SCI) and INSPEC databases for the period from 1980 to January 1996. These searches were conducted by examining abstract summaries of approximately 300 different publications selected from computerized searches using the keyword "bistatic."

Section II of this report serves primarily as a guide to the open literature on bistatic clutter measurements. Several of the measurement programs discussed in this section resulted in the collection of low grazing angle bistatic clutter measurements. Knowledge of such measurements is currently critical for designing future land-based and air-borne bistatic radar systems. Therefore, part III of this paper provides a tabular summary of the low grazing angle bistatic

clutter database, in which the NRCS values are tabulated versus terrain type, frequency, polarization, and bistatic scattering geometry.

This report summarizes the author's work performed over a 30 day period during the summer of 1996 as a participant in the AFOSR summer faculty research program at Rome Laboratory, Hanscom AFB. In accordance with the AFOSR\Research Development Laboratory (RDL) policy, this report should be viewed as a very limited distribution working draft of a more detailed technical report that will be completed and be more widely circulated in the near future.

II. LITERATURE REVIEW

Prior to 1982, the unclassified literature contained only five distinct sets of bistatic clutter measurement results [7-11], as summarized in [5] and [6]. Additional bistatic clutter measurements performed since these review reports were written include the work performed by Georgia Institute of Technology in 1982 [12], the University of Michigan in 1988 [13], the Syracuse Research Corporation during the early-mid 1990's under the RL/Griffis-sponsored AMBIS program [14], Northeastern University under the RL Hanscom-sponsored bistatic test bed experiment program [15], and a limited set of measurements performed by the University of Nebraska [16]. Additional bistatic terrain scattering measurements have also been conducted by MIT-Lincoln Laboratory under the joint Navy/ARPA Mountain Top program, however published results from this program will not be available until the Fall of 1996 [17]. Table I provides a summary of these measurement programs; they are discussed individually below. Additional details concerning several of the experiments described below are given in the companion paper by Twarog [18]

A. Bistatic Clutter Measurements Conducted Prior to 1982.

Cost and Peake of Ohio State University [7] conducted a set of bistatic clutter measurements using a short-range (13 foot) X-band radar system viewing different types of terrain samples installed atop a railroad car bed. Both transmitting and receiving antennas were ground-based, and grazing angles ranged from 5 to 90 degrees (nadir). Six types of terrain of varying degrees of roughness, including sand, loam, grass and soybeans were measured over a

wide range of bistatic angles at three different antenna polarization combinations (VV, HH, and HV). Low grazing angle NRCS values for sand and soybeans ranged from -10 dB to -40 dB for out-of-plane angles between 30 and 140 degrees at VV and HH polarizations. Ten curves describing the variation in terrain clutter NRCS with geometry and polarization are included in Cost and Peake's report [7], and these curves are reproduced in Weiner's report [5]. The radar system used in this experiment was calibrated using an aluminum sphere.

Pidgeon [8] conducted in-plane C- and X-band bistatic sea clutter measurements using a land-based transmitter and an airborne receiving system. Data were collected at transmitter depression (grazing) angles ranging from approximately 1.0 to 3.0 degrees and receiver depression angles ranging from 10 to 90 degrees, for a limited range of sea states. Measurements were conducted at C-Band using a CW radar and at X-Band using a pulsed radar system, the radars operating at VV and HH polarizations, respectively. The measurements have been published in a refereed journal article [8] and data curves are reproduced by Weiner [5]. The radar systems were calibrated using the line-of-sight technique. Additional details of this experiment are contained in the paper by Twarog [18].

Pidgeon's measured NRCS values were observed to generally decrease with decreasing transmitter grazing angle, while remaining relatively independent of receiver grazing angle. In-plane NRCS values ranged between -45 and -55 dB for moderate (10-30 knot ocean surface wind speeds) and between -55 and -65 dB for lower windspeeds, as transmitter grazing angle ranged between three and one degree. The authors observed that cross-polarized NRCS decreased with decreasing transmitter grazing angle at a faster rate than the copolarized NRCS, depolarization ratios (ratio of cross-polarized to co-polarized NRCS) ranging between -5 and -8 dB at a three degree grazing angle and between -10 and -15 dB at one degree transmitter grazing angle. HH-polarized X-band NRCS values were between -30 and -40 dB for this range of grazing angles for a Beaufort 5 characterized sea state.

Goodyear Aerospace Corporation [9] conducted a set of millimeter wave (95 GHz) bistatic clutter experiments during the late 1970's. Using ground-based transmitting and receiving antennas, they measured the NRCS of cotton fields and desert terrain at grazing incidence over out-of-plane scattering angles ranging from 70 to 180 degrees. Backscatter-plane

NRCS values were -20 dB and -35 dB for green cotton covered fields and for desert terrain, respectively. Clutter cross sections were observed to decrease with increasing out-of-plane angle. The NRCS of desert terrain was observed to decrease from -35 dB to -50 dB as the out-of-plane angle was decreased from 180 to 90 degrees. The radar was calibrated using an aluminum sphere.

During the late 1970's, the Environmental Research Institute of Michigan and the University of Michigan investigated bistatic clutter using a short pulse (100 nSEC pulse width) radar system operating at both L and X-Bands and at HH polarization [10]. The radar transmitter was mounted on a fixed tower while the receiver was installed in an aircraft. Data were collected at antenna depression angles (grazing angles) of 5, 10, 15 and 20 degrees and at out-of-plane angles ranging from zero degrees (in-plane) to 180 degrees. For a region of rough terrain (dry tall weeds and scrub trees) and for a second clutter region of dry grass, X-band cross sections were typically -10 dB at both small and large out-of-plane angles (30 and 180 degrees, respectively), but NRCS levels exhibited a 15-25 dB decrease near 90 degrees out-of-plane. L-band cross sections were typically 10 dB below X-band levels, with typical cross sections of -20 dB at small and large out-of-plane angles. Forward scatter NRCS levels close to 0 dB were observed for both terrain types at these frequencies. The radar was calibrated using the line-of-sight technique.

During 1977 and 1979, Raytheon Company [11] collected in-plane bistatic clutter data at X-Band at grazing incidence and HH polarization. They observed forward-scatter cross sections between zero and 20 dB for bare and snow-covered grass.

B. Bistatic Clutter Measurements Conducted After 1982

Ewell and Zehner [12], from the Georgia Institute of Technology Engineering Experiment Station, conducted a limited number of bistatic sea clutter measurements during the early 1980's. They used a high-power (250 kW) X-band radar with truck and tower mounted transmitting antennas and collected HH and VV polarized data at grazing incidence and bistatic angles ranging from five to 120 degrees from the forward-scatter plane. Rather than reporting calibrated NRCS quantities, these investigators reported the ratio of bistatic to monostatic cross

sections. Ratios ranged between -5 dB and -25 dB, decreasing with increasing out-of-plane angle, and with HH-polarized ratios typically being several dB larger than VV-polarized ratios.

During the mid 1980's, the University of Michigan conducted laboratory measurements of bistatic clutter cross section at 35 GHz [13]. They used a network analyzer-based radar and observed terrain samples at short ranges at VV, HH, and HV polarizations. The radar was calibrated using both the line-of-sight technique and a flat plate. Their measurements were conducted at high grazing angles of 30 or 24 degrees (incidence angles of 60 and 66 degrees, respectively), to approximate the brewster angle for smooth sand. Forward scatter NRCS levels of 25 and 13 dB were observed for HH and VV polarizations, respectively. Cross sections were found to initially decrease with increasing out-of-plane angle, reaching minima of -25 dB for HH polarization and -30 dB for VV polarization at out-of-plane angles between 30 and 60 degrees. For out-of-plane angles between 90 and 180 degrees, cross sections ranged between -5 and -15 dB for VV polarization and -10 to -20 dB for HH polarization. Cross-polarized NRCS was found to be substantially below co-polarized NRCS in the forward scatter plane, the depolarization ratio being below -40 dB. Cross-polarized NRCS exceeded co-polarized NRCS by 10 dB at a 45 degree out-of-plane angle.

Syracuse Research Corporation has recently performed bistatic clutter measurements under the Rome Laboratory/Griffis-sponsored AMBIS program [14]. They operated an S-Band radar system composed of a cooperative aerostat-based transmitter and a receiver operated from a small aircraft. They performed bistatic clutter measurements over ocean, Gulf of Mexico, Florida Keys, and Florida everglades terrains. Transmitter and receiver grazing angles ranged between one and 17 degrees, while out-of-plane scattering angles ranged from 20 to 135 degrees. Additional details describing this experiment are summarized in [18].

During 1994 and 1995, Northeastern University collected S-Band bistatic clutter data at low grazing angles and 25-75 degree out-of-plane angles using a high-power transmitter and a dual-polarized receiver, separated from each other over a 20 km baseline [15]. The measured terrain included a region of rolling, forested hills in Eastern MA. The measurements were fully polarimetric, and the radar system was calibrated using a line-of-sight technique. Measured cross

sections ranged between -34 and -56 dB, decreasing with increasing out-of-plane scattering angle. These measurements are discussed further in section III.

Narayanan of the University of Nebraska measured the X-Band bistatic NRCS of individual trees using a short-range network analyzer-based radar system [16]. Their report tabulated radar cross section (RCS) of individual trees rather than NRCS values for distributed foliage canopies. Vertically co-polarized RCS was found to average -2.4 dB while cross-polarized RCS averaged -7.3 dB, over the range of out-of-plane angles between 10 and 90 degrees.

III LOW GRAZING ANGLE MEASUREMENTS

Several of the measurement programs described in section II resulted in low grazing angle bistatic clutter measurements. Knowledge of such data is critical for the design and performance prediction of future airborne and land-based bistatic radar systems. Rough surface scattering models are routinely used to predict bistatic clutter cross sections at high grazing angles (above a few degrees), however such models consistently underpredict bistatic NRCS levels at lower grazing angles, and many computer prediction codes therefore include "floor" values in their NRCS predictions. This section concentrates on low grazing angle (defined for this presentation as grazing angles below five degrees) measurement results that were included in the reports by the Ohio State University, Goodyear Aerospace Corporation, University of Michigan, Syracuse Research Corporation, and Northeastern University. Calibrated NRCS values for a variety of terrain types are included in Table II.

Table I. Summary of Bistatic Clutter Measurement Programs

Reference	Date	Organization	Frequency	Polarization	Terrain	Incidence Angles		Out-of-Plane Angles [deg]
						Tx [deg]	Rx [deg]	
[7]	1965	Ohio State University	X-Band	VV, HH, HV	Soil, Sand, Grass	20-85	0-85	0-180
[8]	1966	JHU/APL	C,X Bands	VV, HV,HH	Ocean	87-89	0-80	180
[9]	1977	Goodyear Aerospace	95 GHz	VV	Fields, Desert	90	90	70-180
[10]	1978	ERIM/U. Michigan	L,X-Bands	HH, HV	Grass, Weeds	50-80	70-85	0-180
[11]	1969	Raytheon Company	X-Band	VV,HH	Grass, Trees	90	90	0
[12]	1982	Georgia Tech	X-Band	VV, HH	Ocean	90	90	120-160
[13]	1988	University of Michigan	35 GHz	VV, HH, HV	Sand and Gravel	60	0-80	0-180
[14]	1995	Syracuse Research Corp.	S-Band	VV	Ocean, Everglades	73-89	73-85	35-135
[9]	1995	Northeastern University	S-Band	Polarimetric	Forested Hills	0-1	0-1	20-75
[10]	1993	University of Nebraska	X-Band	VV,VH	Individual Trees	90	90	10-90

Table II. Summary of Low Grazing Angle NRCS Values

Investigator	Terrain Type	NRCS [dB]	OOP Angle	Polarization	Frequency
OSU	Sand	-30 - -40	30 - 140	HH	X-Band
	Sand	-30 - -45	30 - 140	VV	X-Band
	Loam	-18 - -30	30 - 140	HH	X-Band
	Soybeans	-30 - -40	30 - 140	HH	X-Band
JHU/APL	Ocean	-45 - -65	in plane	VV	C-Band
	Ocean	-30 - -40	in plane	HH	X-Band
Goodyear	Desert	-38 - -48	25 - 110	VV	95 GHz
U. Michigan	Dry Grass	-10 - -15	30 - 120	HH	X-Band
	Dry Grass	-15 - -25	30 - 120	HH	L-Band
SRC	Ocean	-45	35 - 135	VV	S-Band
	Everglades	-30 - -35	75 - 125	VV	S-Band
	Keys/Mixed	-35 - -45	55 - 115	VV	S-Band
Northeastern	Trees/Hills	-34 - -56	25 - 75	VV,HH	S-Band

With the exception of the Johns Hopkins University (JHU/APL) data set, all NRCS values in this table describe out-of-plane scattering at low grazing angles. The JHU/APL, Goodyear, and Northeastern University measurements were collected at extremely low grazing angles (nearly 90 degrees incidence), whereas the other datasets were obtained at grazing angles of several degrees. The University of Michigan data were collected at a five degree transmitter grazing angle but at a higher receiver grazing angle. The NRCS levels are specified over a range of values owing to the change in the incidence angle and also to changes in the out-of-plane angle for a particular experiment, frequency band, polarization, and terrain type. In general, NRCS levels decrease with decreasing incidence angle. The variation in NRCS with out-of-plane angle is non-monotonic; NRCS typically decreases with increasing angle away from the forward-scatter plane, reaches a minimum near 90 degrees, and then increases with increasing out-of-plane angle. This behavior is evident in most of the datasets indicated.

Due to the limited number of measurements acquired, it is difficult to intercompare NRCS measurement results among experiments. It can be seen from Table II, however, that in general, X-band NRCS levels are larger than S-band levels. For low grazing angles and out-of-plane angles several tens of degrees from the forward scatter plane, S-band NRCS values range from -30 dB to -56 dB for terrain regions that exclude water. X-band NRCS levels range from -18 dB to -40 dB for similar geometries and terrain types. Thus, it can be concluded, with caution, that X-band bistatic clutter levels are 10-15 dB higher than S-band levels.

IV. CONCLUSIONS

The ten bistatic clutter measurement programs documented in the open literature have been identified and briefly summarized in this report. Low grazing angle NRCS values resulting from these experiments have been tabulated. S-band values range from -34 dB to -56 dB, and X-band NRCS levels are 10 to 15 dB higher. Additional bistatic clutter measurements are needed to validate and extend this database. The measurements of Cost and Peake as well as the 35 GHz measurements performed by the University of Michigan are laboratory-based, and full-sized field measurements are needed for validation of these results. C-band bistatic data are virtually nonexistent; the JHU/APL dataset it limited to in plane data at a few different sea states, and no

terrain measurements have been performed in this frequency band. S- and X-band data are best represented in the existing database, but many different terrain types and geometries must be studied to improve upon the present understanding of bistatic clutter phenomenology.

V. REFERENCES

[1] Y.S. Hsu and D.C. Lorti, "Spaceborne Bistatic Radar - An Overview," IEE Proc. F., 1986, Vol. 133, No. 7, pp. 642 - 648.

[2] J. Wurman, M. Randall, C. Frush, E. Loew, and C.L. Holloway, "Design of Dual Doppler Radar for Retrieving Vector Winds using One Transmitter and a Remote Low-Gain Receiver," Proc. IEEE, 1994, Vol. 82, No. 12, pp. 1861 - 1872.

[3] R.L. Fante, "Cancellation of Specular and Diffuse Jammer Multipath using Hybrid Adaptive Arrays, " IEEE Trans. AES-27, 1991, No. 5, pp. 805 - 820.

[4] R.L. Fante and J.A. Torres, "Cancellation of Diffuse Jammer Multipath by an Airborne Adaptive Radar," IEEE Trans. AES-31, 1995, No. 2, pp. 805 - 820.

[5] M.W. Weiner, "Report on Bistatic Scattering Phenomenology," MIRTE Corporation Technical Report, 1981.

[6] R.J. Papa and J.F. Lennon, "L-Band and S-Band Land and Sea Clutter Models," RADR-TR-81-351, 1981.

[7] W.H. Peake and T.L. Oliver (1971), "The Response of Terrestrial Surfaces at Microwave Frequencies," Electronics Research Laboratory Technical Report, Ohio State University, TR-2440-7, May 1971.

[8] V.W. Pidgeon, "Bistatic Cross Section of the Sea," IEEE Transactions on Antennas and Propagation, Vol. AP-14, No. 5, May, 1966, pp. 405 - 406.

[9] L.P. Johnson and V.M. Fey, "Bistatic Data Collection," Goodyear Aerospace Corporation, Litchfield Park, Arizona, Technical Report Number AFAL-TR-77-274.

[10] R.W. Larson, A.L. Maffett, R.C. Heimiller, A.F. From, E.L. Johansen, R.F. Rawson, and F.L. Smith, "Bistatic Clutter Measurements," IEEE Transactions on Antennas and Propagation, Vol. AP-26, No. 6, November 1978, pp. 801 - 804.

[11] P.E. Cornwall, A.H. Greene, and D.G. Armstrong, "Multipath Measurements, Phase Three," Raytheon Company, Equipment Development Laboratory, Final Report Number ER-77-4121.

[12] G.W. Ewell and S.P. Zehner, "Bistatic Sea Clutter Return Near Grazing Incidence," IEE Conference Publication 216 (Radar 82), 1982, pp. 188 - 189.

[13] F.T. Ulaby, T.E. Van Deventer, J.R. East, T.F. Haddock, and M.E. Coluzzi, "Millimeter-Wave Bistatic Scattering from Ground and Vegetation Targets," IEEE Transactions on Geoscience and Remote Sensing, Vol. 26, No. 3, May 1988, pp. 229 - 243.

[14] Syracuse Research Corporation, "Bistatic Clutter Measurement Program, Phase 3," Contract Number F30602-93-D-0081, SRC-TR-95-180, May 1995.

[15] D.J. McLaughlin, R.S. Raghavan, W.G. Stevens, and M.J. Sowa, "Bistatic Terrain Clutter Dependence on Out-of-Plane Scattering Angle," Electronics Letters, July 20, 1995, Vol. 31, No. 15, pp. 1291 - 1292.

[16] R. M. Narayanan, S.E. Nelson, and J.P. Dalton, "Azimuth Scattering Pattern of Trees at X-Band," IEEE Trans. AES-29, 1993, No. 2, pp. 588-593.

[17] R. Gabel, MIT-Lincoln Laboratory, Private Communication

[18] E.M. Twarog, "Airborne Bistatic Clutter Measurements: Systems Issues", this issue.

NEURAL BEAM STEERING AND DIRECTION FINDING

H. N. Mhaskar
Professor
Department of Mathematics
California State University
Los Angeles, CA 90032

Final Report for:
Summer Faculty Research Program
Rome Laboratories/ERAA
Hanscom AFB, MA 01781

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC
and
Rome Laboratories/ERAA

July, 1996

NEURAL BEAM STEERING AND DIRECTION FINDING

H. N. Mhaskar
Department of Mathematics
California State University
Los Angeles, CA 90032

Abstract

In this paper, we investigate algorithms for direction finding and beam steering using a degraded antenna. The algorithms can be implemented by means of neural networks. The training of these neural networks does not involve any nonlinear optimization. On several data sets, our results are comparable with those obtained with previously studied radial basis function networks. In the case of direction finding, they provide a substantial improvement.

NEURAL BEAM STEERING AND DIRECTION FINDING

H. N. Mhaskar
Department of Mathematics
California State University
Los Angeles, CA 90032

1 Introduction

Let $m \geq 2$ be an integer. A phased array antenna with m *elements* may be thought of as a function $\boldsymbol{\phi} : [-\pi/2, \pi/2] \to [0, 2\pi)^m$. For a plane wave incident on the antenna at an angle θ in $[-\pi/2, \pi/2]$, $\boldsymbol{\phi}(\theta)$ represents *measured phases* at each of the elements. In the ideal circumstances, the function $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_m)$ is simply given by

$$\phi_k(\theta) = \psi(\theta) + k\mu \sin\theta, \qquad k = 1, \cdots, m, \tag{1}$$

where ψ is some function, and μ is a constant depending upon the architecture of the antenna. The equation (1) can be used to find the angle θ, given $\boldsymbol{\phi}(\theta)$. Indeed, one has

$$\theta = \arcsin((\phi_{k+1}(\theta) - \phi_k(\theta))/\mu), \qquad k = 1, \cdots, m-1. \tag{2}$$

Since this equation is independent of ψ, it is customary to assume (and force in calculations) that $\psi(\theta) \equiv 0$.

When the antenna is degraded, then (1) cannot be expected to hold. Nevertheless, it is reasonable to suspect that the (unknown) function $\boldsymbol{\phi}$ continues to be a one-to-one function. In practice, the useful domain of $\boldsymbol{\phi}$ is typically $[-\pi/3, \pi/3]$ rather than $[-\pi/2, \pi/2]$, and even smaller, as we will observe later. The problem of *beam steering* consists of approximating the function $\boldsymbol{\phi}$. The problem of *direction finding* consists of approximating an inverse function of $\boldsymbol{\phi}$. This approximation is based on a series of experiments where the values of the function are observed for a range of values

of incidence angles θ. It is now evident that neural networks provide a natural paradigm for solving these problems, both because of their ability to approximate unknown functions, and because of their parallel processing capabilities.

Several such experiments were performed at the Rome Laboratories, Hanscom AFB by a group lead by Dr. H. Southall, resulting in several data sets. In this paper, we have worked with five of these: Test97, Test87, Test92, Test60, and Testsc. We have also worked with a data set, Tesh22, constructed artificially from Test97 by disabling two of the elements of an eight element array. In [3, 4], radial basis function networks have been trained to learn the unknown function ϕ , as well as its inverse based on these data sets. Each of these data sets consists of the observations of ϕ at 1° interval in $[-60°, 60°]$, 121 readings altogether. Typically, the "training data" consists of the observations $\{\phi_k := \phi(-\pi/3 + (k-1)\pi/18)\}_{k=1}^{13}$. The resulting networks evaluate a function of the form

$$\sum_{k=1}^{13} w_k \exp(-||\cdot - \phi'_k||^2/\sigma^2) \tag{3}$$

where $\{\phi'_k\}$ are the "phase differences" obtained from $\{\phi_k\}$ after some adjustments, and the training consists of determining the correct values of w_k's and σ. Clearly, the hardest part is the determination of σ; which is actually done on the basis of the entire data set of 121 readings.

During the first-named author's visit to the Rome Laboratories from May 20, 1996 to July 28, 1996, we explored alternative activation functions for the neural networks. It is shown in this paper that the activation function

$$\Phi(x) := \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

can be used effectively for both the problems, along with some preprocessing and postprocessing of the data. The approximation capabilities of networks using this activation function are studied in [2, 1].

The most attractive feature of our algorithms is that they involve no nonlinear optimization, similar to the one used to obtain the value of σ in the Gaussian networks described in [3, 4]. In particular, they represent "genuine" training on samples with no reference to the whole data set.

2 Beam Steering

The idea behind our algorithm is the following. As in [3, 4], our starting point is the function $\mathbf{x} : [-\pi/3, \pi/3] \to [0, 2\pi)^{m-1}$, where (with $\mathbf{x} = (x_1, \cdots, x_{m-1})$),

$$x_k(\theta) := \phi_{k+1}(\theta) - \phi_k(\theta), \qquad k = 1, \cdots, m-1. \tag{5}$$

It was thought reasonable to suspect that in spite of the degradation of the antenna, the function $\mathbf{x}$ would still have the form

$$x_k(\theta) = \mu \sin\theta + \epsilon_k(\theta), \qquad k = 1, \cdots, m-1 \tag{6}$$

where $\epsilon_k(\theta)$ represents a "noise" with mean zero. Therefore, all coordinates of $\mathbf{x}$ are essentially equal, and may be replaced by their average. A straightforward average, however, amounts to neglecting all but the first and the last element. Therefore, we seek to approximate the x_k's by

$$\hat{x}_k(\theta) = \sum_{j=1}^{m-1} a_j x_j(\theta) = \mathbf{a} \cdot \mathbf{x}(\theta), \qquad k = 1, \cdots, m-1, \tag{7}$$

for a judicious choice of $\mathbf{a} \in \mathbf{R}^{m-1}$. We observe again that $\hat{x}_k$ is really independent of k; it represents the "real" phase difference hidden in the observed phase differences in the degraded antenna.

There is one major difficulty with this idea. The phases ϕ_k are observed only modulo 2π. The unwrap function in matlab allows one to remove the resulting branch cuts. However, this function utilizes the entire data set. Therefore, for a proper beam steering algorithm, one needs to approximate the phases (as unwrapped with the unwrap function of matlab) based only on the training data. It is not necessary to try to approximate each phase difference individually; we are looking only for a constant "representative" for each phase difference.

In the algorithm in Figure 3, we solve this problem as follows. We assume that all the desired angles belong to a finite set S, and the training data is of the form $\{\theta_i, \mathbf{x}(\theta_i)\}_{i=1}^n$, where $\theta_1 < \cdots < \theta_n$ are angles from S, measured in radians, and for any θ, $\mathbf{x}(\theta) \in [0, 2\pi)^{m-1}$ denotes the vector of observed phase differences $\{\theta_i, \mathbf{x}(\theta_i)\}_{i=1}^n$, where $\theta_1 < \cdots < \theta_n$ are angles from S, measured in radians. We assume that

$$\theta_1 = \min S, \qquad \theta_n = \max S. \tag{8}$$

We start by removing the branch cuts only in the training data, using the unwrap function. For the intermediate angles, we just estimate the desired phase difference by a piecewise linear interpolation of the unwrapped training phase differences, thereby obtaining a model for the whole data set. In Figure 1 and Figure 2, we show the third phase difference for the data sets Test92 and Tesh22 obtained as a result of this interpolation as well as that obtained by "unwrapping" the whole data sets using the unwrap command of matlab. Some minor adjustments at the angle $\theta = 0$ are made so as to ensure that the resulting phases agree with the observations at this angle. We also use the training data to obtain a good vector $\mathbf{a}$ that works in (7), and then obtain the desired "representative" phase differences by using the model data set. A further refinement is obtained using the entire model by using a linear regression of $\mathbf{y}$ on the quantities $\mu \sin\theta$.

In Figure 3, the first step is the preprocessing of data, steps 2, 3, 4 train a neural network with $m-1$ outputs, and $m-1$ neurons, each evaluating the activation function given by (4). The last three steps constitute the postprocessing. We observe that the output of the algorithm is a scalar, notwithstanding the notation.

We may use two techniques to test the performance of our algorithm. Both of these are based on the fact that the maximum absolute value of $\sum_{k=1}^{m-1} \exp(ik\theta)$ is assumed when $\theta = 0$ (modulo 2π). Thus, having obtained the predicted constant phase difference $\hat{\mathbf{x}}$, we may find the actual direction in which the beam is formed by the expression

$$\arg\max_{\theta \in S} \sum_{k=1}^{m-1} \exp(ik(\mu \sin\theta - \hat{\mathbf{x}}))$$

We call this "Method 2". Alternatively, we may argue that for the degraded array, one should take the measured phases for the desired direction as the ones needed to produce the beam in that direction. With this viewpoint, it more reasonable to consider the value of $\mathbf{y}$ obtained in Step 4 as the final prediction of the algorithm, and calculate the expected beam direction by the expression (cf. (5))

$$\arg\max_{\theta \in S} \sum_{k=1}^{m-1} \exp(i(\sum_{j=1}^{k}(x_j(\theta) - y_j))). \tag{13}$$

We call this "Method 1". The Table 1 summarizes the RMS error in degrees for both these methods, as well as for the method using RBF networks, as

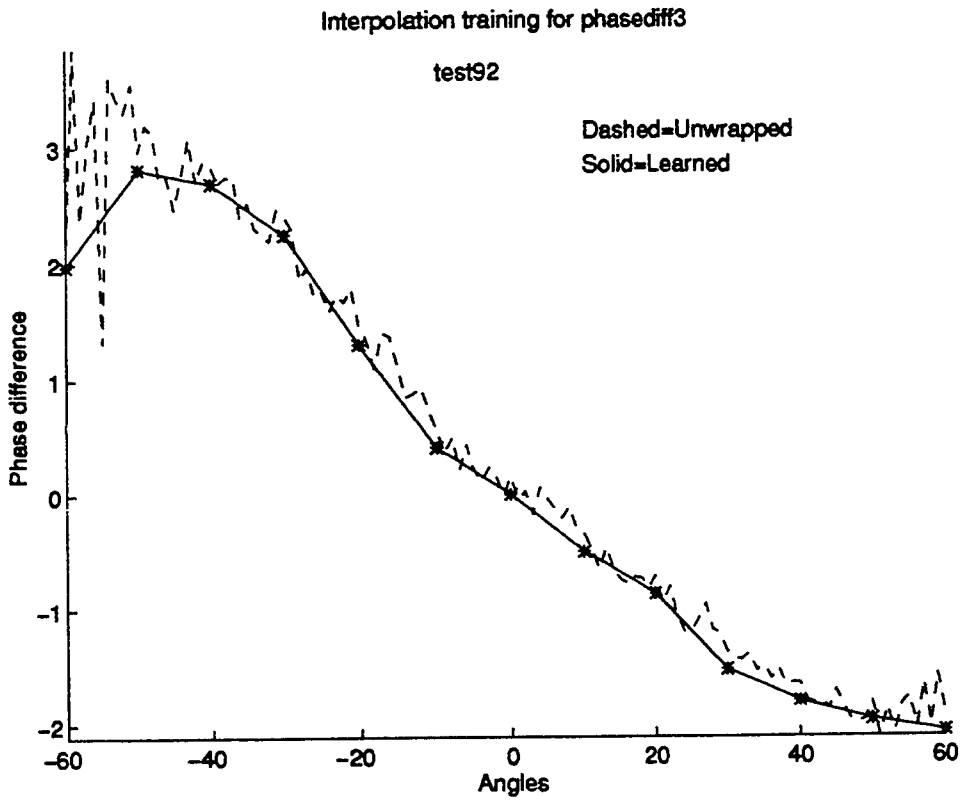


Figure 1: Phase differences 3 for Test92, obtained by interpolation and by the unwrap command of matlab used on the entire data set.

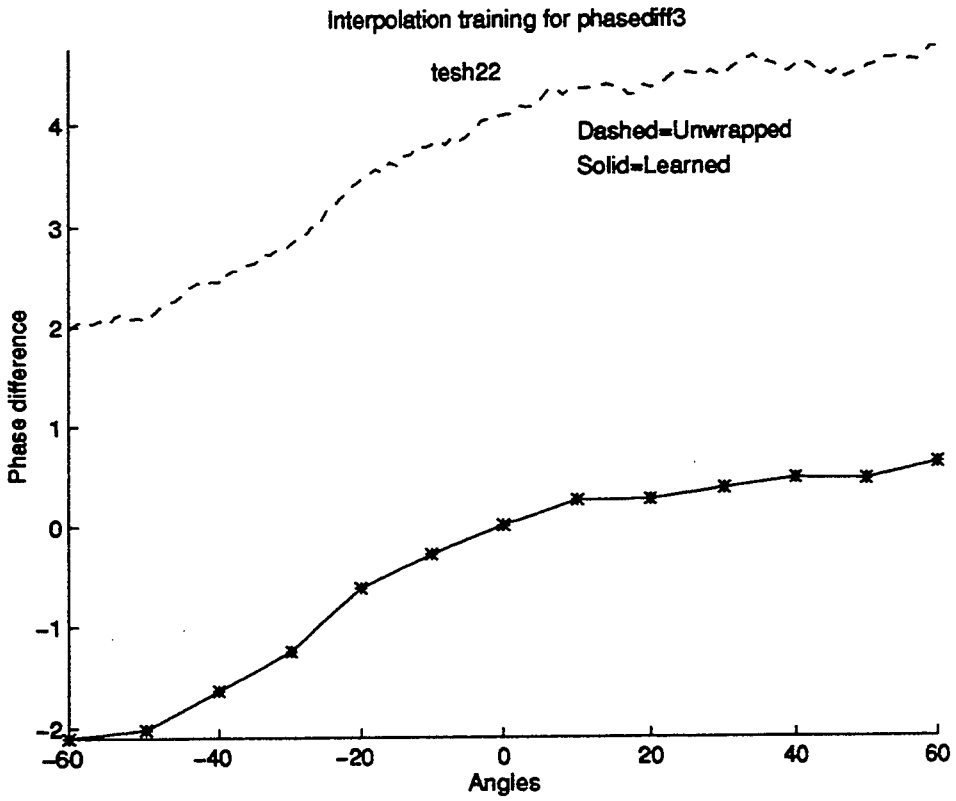


Figure 2: Phase differences 3 for Tesh22, obtained by interpolation and by the unwrap command of matlab used on the entire data set.

1. Remove branch cuts from the training data, obtaining a new training data $\{\theta_i, \mathbf{y}(\theta_i)\}$.

2. (Linear interpolation)

   For each $\theta \in S$, do
   Find j such that $\theta_j \leq \theta < \theta_{j+1}$, and set

$$\mathbf{y}(\theta) := \frac{(\theta - \theta_j)\mathbf{y}_{j+1} + (\theta_{j+1} - \theta)\mathbf{y}_j}{\theta_{j+1} - \theta_j} \tag{9}$$

3. Write $\mathbf{d} := \mathbf{x}(0) - \mathbf{y}(0)$.

4. For each $k = 1, \cdots, n$, do

$$\mathbf{y}(\theta_k) := \mathbf{y}(\theta_k) + \mathbf{d}.$$

5. Find $\mathbf{a}^* \in \mathbf{R}^{m-1}$ by

$$\mathbf{a}^* := \arg \max_{\mathbf{a} \in \mathbf{R}^{m-1}} \sum_{k=1}^{m-1} \Big(\mu \sin \theta_k - \mathbf{a} \cdot \mathbf{y}(\theta_k) \Big)^2. \tag{10}$$

6. Find $(A^*, B^*) \in \mathbf{R}^2$ by

$$(A^*, B^*) := \arg \min_{(A,B) \in \mathbf{R}^2} \sum_{\theta \in S} \Big(\mu \sin \theta - A\mathbf{a}^* \cdot \mathbf{y}(\theta) - B \Big)^2. \tag{11}$$

7. Return

$$\hat{\mathbf{x}} = A^*(\mathbf{a}^* \cdot \mathbf{y}(\theta)) + B^*. \tag{12}$$

Figure 3: Beam Steering algorithm, generates a phase difference $\hat{\mathbf{x}}$ given a desired beam direction θ.

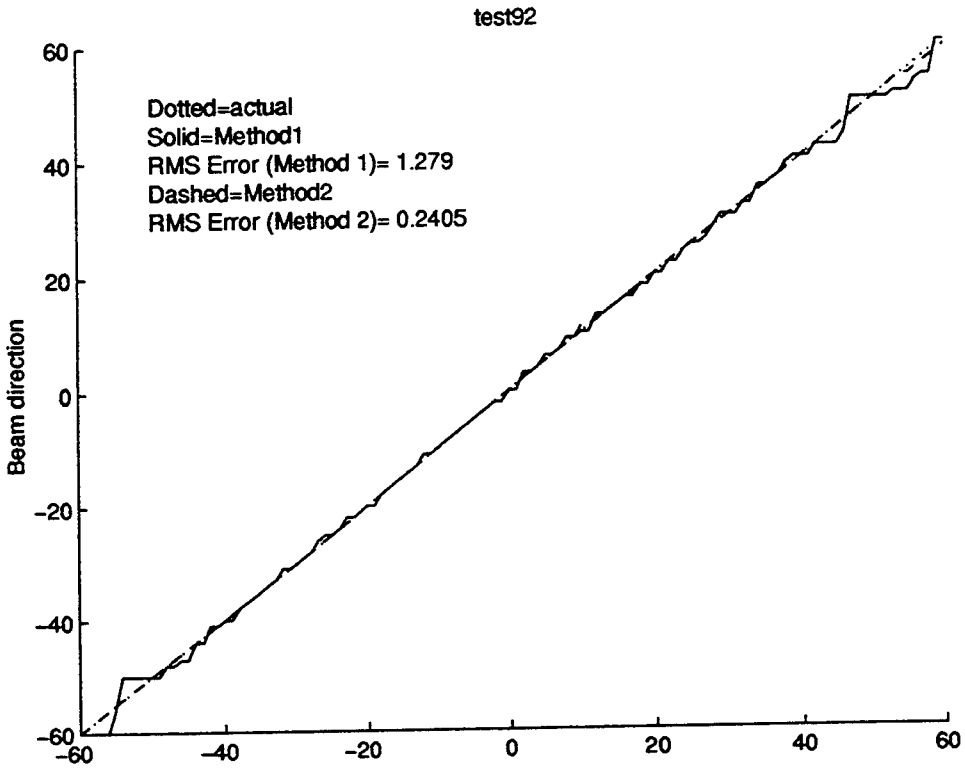


Figure 4: Beam Steering for Test92

applied to the six data sets, five of which are described in [3]. The results for Test92 and Test60 are shown graphically in Figures 4 and 5 respectively.

3 Direction finding

The problem of direction finding is more difficult than that of beam steering, mostly because one wishes to approximate the inverse function of ϕ. Although it is reasonable to expect that the inverse function exists, even for a degraded antenna, the periodicity of the observations make it very difficult to determine the point at which the inverse function should be evaluated! Surprizingly, the idea behind the Method 1, used to test our beam steering experiments, gives a good solution to this problem. Our simple algorithm for direction finding is given in the following Figure 6. We note that the steps 1

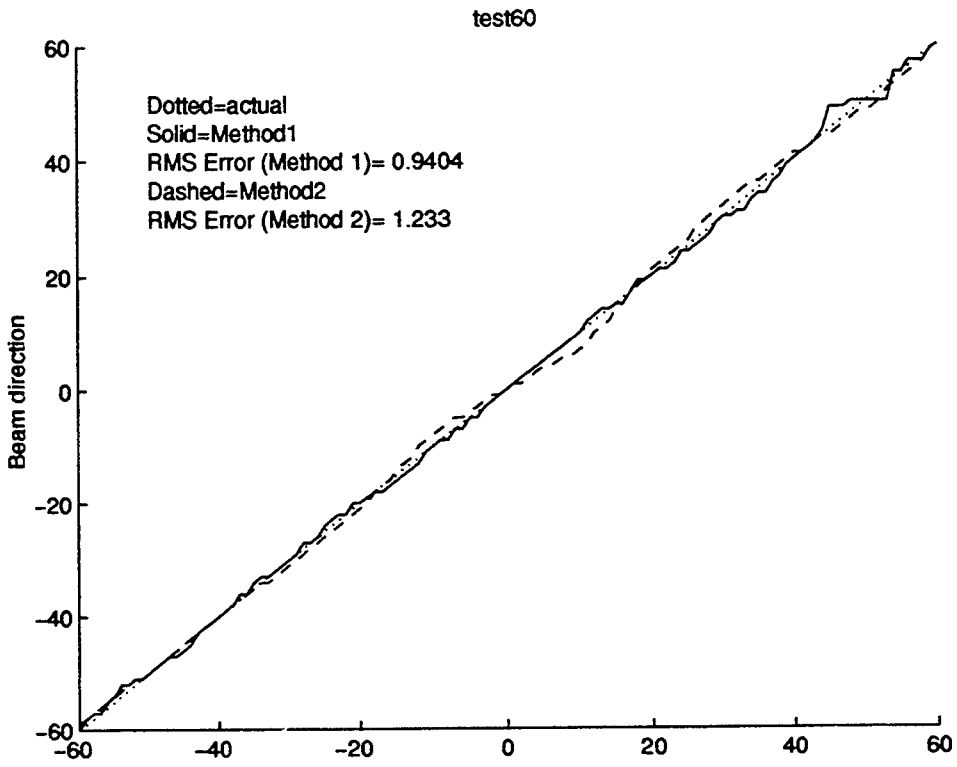


Figure 5: Beam Steering for Test60

Dataset	RBF	Method 1	Method 2
Test97	0.34	0.18	0.24
Test87	0.48	0.31	0.09
Test92	1.24	1.28	0.24
Test60	0.94	0.94	1.23
Testsc	1.05	0.92	0.76
Tesh22	0.55	0.22	0.22

Table 1: RMS errors (degrees) in beam steering experiments

Dataset	RBF	Our method
Test97	0.41	0.16
Test87	1.41	0.48
Test92	1.97	1.39
Test60	1.93	1.13
Testsc	1.59	1.10
Tesh22	0.59	0.09

Table 2: RMS errors (degrees) in direction finding experiments

and 2 are the same as in Figure 3. Our assumptions are also the same as in that algorithm. Once more, Step 1 is the preprocessing part. Steps 2 and 3 constitute neural network training, and Steps 4 and 5 are the postprocessing part of the algorithm.

We observe that the calculation of the angle in Step 5 in the algorithm of Figure 6 is slightly different from that in testing the beam steering algorithm using (13). In testing an algorithm, it is legitimate to use the entire data set as in (13). However, in the direction finding algorithm, we do not have an access to the whole data set; we have to work only with the one observation of the phases and the training data set. Therefore, in Step 5 of the direction finding algorithm, we use the model constructed from the training data. Consequently, our results are not exactly the same as in Table 1. The RMS errors in degrees for each of the data sets in Table 1 are given in Table 2. In [3], many different algorithms using RBF networks are discussed. In Table 2, we have taken the best result for each data set. The actual training algorithm for the RBF networks is different for different data sets. Nevertheless, it is seen readily that we have realized a substantial improvement over the results obtained by using the RBF networks in each case. Figures 7 and 8 show the results graphically for Test92 and Test60 respectively.

References

[1] H. N. Mhaskar,*Approximation properties of a multilayerd feedforward artificial neural network,* Advances in Computational Mathematics, **1**

1. Remove branch cuts from the training data, obtaining a new training data $\{\theta_i, \mathbf{y}(\theta_i)\}$.

2. (Linear interpolation)

   For each $\theta \in S$, do
   Find j such that $\theta_j \leq \theta < \theta_{j+1}$, and set

$$\mathbf{y}(\theta) := \frac{(\theta - \theta_j)\mathbf{y}_{j+1} + (\theta_{j+1} - \theta)\mathbf{y}_j}{\theta_{j+1} - \theta_j} \tag{14}$$

3. For each $\theta \in S$, construct the phases $\mathbf{P}(\theta) = (P_1, \cdots, P_m)$ by

$$P_j = \begin{cases} 0, & \text{if } j = 0, \\ \sum_{k=1}^{j} y_j(\theta), & \text{if } 2 \leq j \leq m. \end{cases}$$

4. Construct $\mathbf{p} = (p_1, \cdots, p_m)$ by

$$p_j = \begin{cases} 0, & \text{if } j = 0, \\ \sum_{k=1}^{j} x_j, & \text{if } 2 \leq j \leq m. \end{cases}$$

5. Return

$$\hat{\theta} := \arg\max_{\theta \in S} \left| \sum_{k=1}^{m} \exp(i(P_k(\theta) - p_k)) \right|.$$

Figure 6: Direction finding algorithm, returns the expected incidence angle $\hat{\theta}$ corresponding to an observed phase vector.

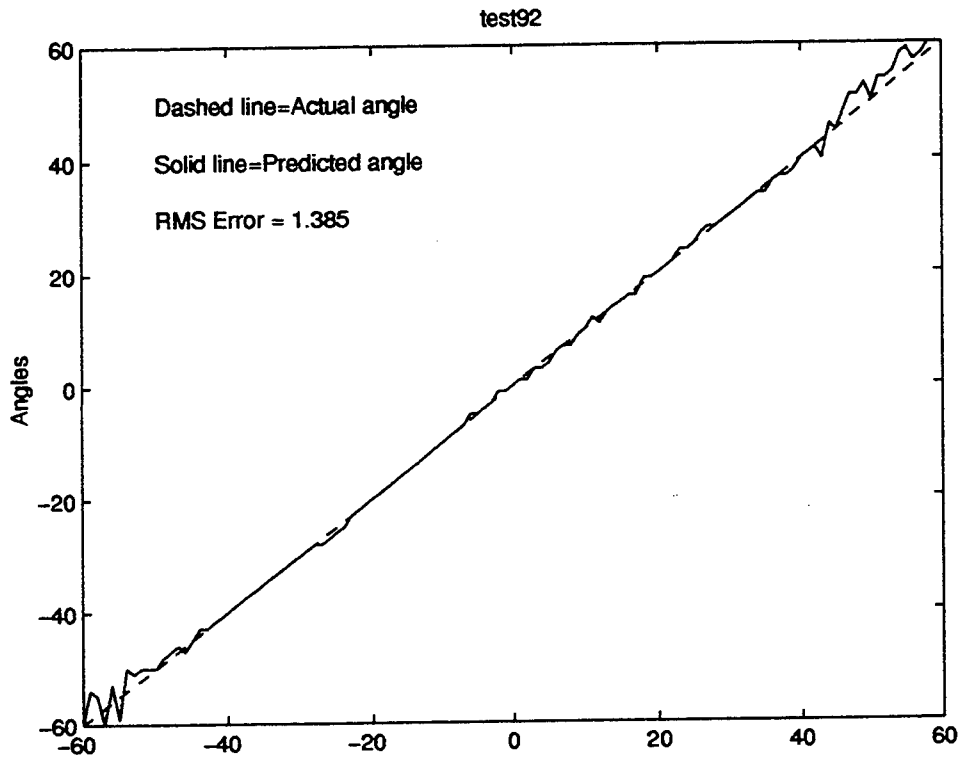


Figure 7: Direction finding for Test92

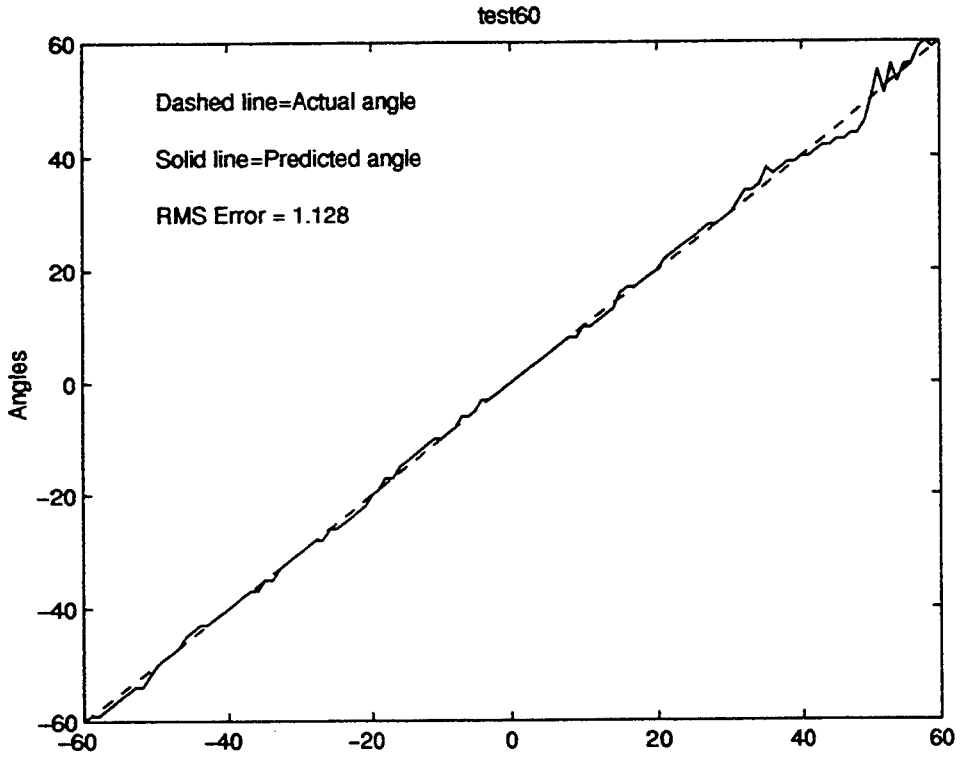


Figure 8: Direction finding for Test60

(1993), 61-80.

[2] H. N. Mhaskar and C. A. Micchelli,*Approximation by superposition of a sigmoidal function and radial basis functions,* Advances in Applied Mathematics, **13** (1992), 350-373.

[3] H. Southall, J. A. Simmers, and T. H. O'Donnell, *Direction finding in phased arrays with a neural network beamformer,* IEEE Transactions on Antennas and Propagation, **43** (12) (1995), 1369–1374.

[4] H. Southall, S. Santarelli, E. Martin, and T. O'Donnell, *Neural beam-steering for phased array antennas,* Proc. of the 1995 Antenna Applications Symposium, University of Illinois, Sept. 1995, USAF Rome Laboratory, Rome, NY.

A LOW DIMENSIONAL CATEGORIZATION TECHNIQUE FOR C SOURCE CODE

Ronald W. Noel
Assistant Professor
Department of Philosophy, Psychology,
and Cognitive Science

Rensselaer Polytechnic Institute
110 8th Street
Troy, NY 12180-3590

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Griffiss Air Force Base, NY

and

Rome Laboratory

October 1996

A LOW DIMENSIONAL CATEGORIZATION TECHNIQUE FOR C SOURCE CODE

Ronald W. Noel
Assistant Professor
Department of Philosophy, Psychology,
and Cognitive Science
Rensselaer Polytechnic Institute

Abstract

The rapid development of efficient reliable software, especially large software, is not a task for which humans are particularly well suited. Humans have difficulty thinking about and manipulating large numbers of interacting factors when they develop complex software. On the other hand, complete automation of the software development task is unlikely since it requires a broad expertise in programming, software design, and an awareness of the goals and intentions that is at present uniquely human. The present paper explores how two new cognitive science techniques in recognition and categorizing can aid humans in the search of legacy software in software engineering. One technique recognizes objects in a vastly multi-dimensional space using eigenspaces, and the other technique automatically processes text using *n*-grams. The paper demonstrates how applying these techniques as cognitive filters might aid in the efficient search of legacy C software source code. A search or browser system that contained such a filter would immediately benefit human programmers reviewing legacy software and would be an important step towards developing an automated system.

A LOW DIMENSIONAL CATEGORIZATION TECHNIQUE FOR C SOURCE CODE

Ronald W. Noel

INTRODUCTION

Goals and Motivation

Software engineering exists in the throes of a cognitive science dilemma. On one hand, the engineering of software is a very human process. It requires broad expertise in programming, software design, and the awareness of the goals and intentions of the software system. For the most part, these human processes are poorly understood by cognitive scientists and have not succumbed to software automatization techniques. The inability to automate such tasks has left software engineering to humans. On the other hand, the creation of software, especially large software, is not a task for which humans are well suited. The need to think in exact detail and to consider a large number of simultaneous interacting factors limit the scope, the size and the reliability of the software that a human or group of humans can produce.

The most reasoned approach to this dilemma is the judicious division of tasks between humans and computers. In such an approach, humans would work in broad terms suited to their abilities to manage the goals and intentions of a system and to specify the system in terms of actions and purposes. The specifications would then be turned into code by the computer using formal methods. One would expect that the creation of such code would be of a much higher quality than anything that humans could produce unaided. Such automated coding from a computer based design management system is the approach that describes much of the work and success of knowledge based system engineering [8, 9].

Further success in raising the quality of software will rely on the computer taking over more and more of the human tasks. The first tasks to consider are those that comprise the better understood cognitive tasks, especially mundane. Automating such tasks would free the humans to work on other tasks. The benefits to software engineering would include greater job satisfaction for the human, a reduction in development time, and an increase in consistency (i.e., software quality). The level of performance might be lower than that of the human expert for the same task, but the benefits could be obtained as long as the software's performance was near the average for human performance. Additionally, because automated software capabilities are independent of humans, they would not be lost through attrition. Instead, they could be developed as separate components and studied and improved towards long term system evolution.

The work of this paper is to investigate such a task. The task is recognizing whether legacy software exists that performs a function that will meet the functional specifications of a software. This system is considered to be a type of cognitive filter that selects and prioritizes legacy software for more detailed analysis by humans or formal methods of software analysis [10]. This work on categorizing a class of complex representations grew out of two desires. One desire was to increase understanding of the interface between holistic (i.e., featureless, non-decomposable) representations created by human experts and machine processing. The other was to apply cognitive science techniques used in natural language processing and visual pattern recognition to the domain of machine program understanding. Uses for a categoriser would be for a filter in a browser, a prefilter that prioritizes software for an automated recovery system, a tool that aides in the segmentation of software into meaningful chunks, and a method for determining the intentions of a software artifact.

When faced with the task of reviewing a vast amount of software, a human would first make a series of quick initial judgments based on holistic pattern recognition to eliminate as many candidates for use as possible, and prioritize any software kept. The remaining software would have to go through some detailed analysis to determine if the software was suitable. Software that passes the quick judgment relies on human holistic pattern recognition ability. Since these are the capabilities that the filter are to emulate a brief discussion of human pattern recognition will be valuable.

Categorization

The semantic categorization of objects is studied under many banners. Each banner represents either a domain of objects to be classified (i.e., images versus text), process (i.e. feature analytic versus template matching), or representation (i.e. neural networks versus propositional), or any combination of the three. Considering such broad topics as data modeling, object recognition, and text comprehension one finds that a dominant approach is the decomposing of complex representations into a small number of components or features with a limited number of relationships. The features and their relationships are then captured in a parameterized model that can easily be manipulated or "understood" by algorithms for computer use. For our purposes, a feature-based decomposition method for developing a filter would be to select a set of key words and phases and a set of rules that would categorize based on the presence and absence of the words and phases in a given text.

Problems occur with such an approach since it is open ended and requires either a human analyst to hand craft the solutions or tailor automated approaches to pick features and discover rules for relationships. The difficulty of this is that it is usually not clear what, if any, set of features will decompose the representation coherently. A representation that cannot be decomposed into a limited set of features and relationships is said to be holistic. Such holistic representations, however, do succumb to atomic level decomposition, and

usually it is the atomic symbols that are manipulated on a machine. I suggest that software source code which is created by humans is one of these non-decomposable representations when in comes to semantic classification or categorization.

Human-computer interaction commonly uses atomic representation of objects, as shown in the following examples: (1) A word processor stores and manipulates the graphemes or visual symbols of natural language. Through the manipulation of the graphemes, the program enables humans to represent human thought in text. (2) Graphics programs create, manipulate, and store pixel definition of points of light. Though the manipulation of the pixels, the program enables humans to represent images and visual perceptions. (3) Through the manipulation of program primitives, a programmer can tailor a machine actions to achieve some task or intention. In each case, the machine interacts with the human at an atomic level of representation and has no semantic or meaningful representation. The task falls to the human to organize the representations though an abstract understanding of the object and the methods of structuring the atomic elements to represent the objects.

In the programming example, the atomic level of representation in the interaction limits the aid the application can offer the user. The human-computer interface could be improved if computers had a meaningful representation of programs. The machine could be used for source code data mining or browsing the internet. It could suggest subroutines to accomplish the tasks needed in a program, determine if the program is fulfilling the programmer's intentions, and determine if the programmer intentions are in-line with higher objectives. It could offer help by locating similar programs and determine if the program corresponds with the programmer's intentions.

Holism

A fundamental schism exists between human intellectual abilities and machine processing capacities. Humans are open systems which tend to act in very holistic and non-deterministic ways. On the other hand, machines are closed systems better suited for analysis in a limited domain. Humans lack tremendous numerical computational speed; yet they can process information holistically in an automatic, rapid, and natural manner. Machines possess tremendous computational capabilities; yet no algorithm exists to perform holistic processes as well as humans do. Typically, these differences have led to an antagonistic interface between humans and computers, however, these same differences could lead to a beneficial synergism. Ideally, a good interactive system would integrate the best human qualities with machine computational capabilities enabling it to outperform either of the two cognitive components alone [7].

The theory regarding holistic processing can be separated into stronger and weaker stances. Under the weaker stance, features may interact with each other through configural processes to form emergent properties or "second-order relational features". Under the stronger stance, the process is completely holistic; that is, its representation is non-decomposable in that no explicit description of features or parts exists outside the context of the object. These stances provide two ways to approach the development of systems to support the holistic representation: The traditional approach of a system which extracts features and manipulates context-free features towards configuration [11], or a system which develops the endeavors to understanding the configuration of the image first, followed by more detailed development of features within the established context [12].

A well-known area in which cognitive researchers study holistic processes is the recognition of objects and, in particular, faces. Theories regarding the recognition of objects and faces have traditionally been distinguished by different perceptual encoding

and representational processes; however, the functional separation of these processes under all conditions of object recognition remains unclear [11]. Much of basic object recognition theory has been based on the decomposition of parts and the analysis of edge features [12, 13, 14]. On the other hand, face recognition theory has been based on more holistic processes which utilize surface characteristics such as texture, color, and shading [15]. Some research suggests that the distinctions between object and face recognition begin to fade when one examines the object recognition processes of experts, who may utilize holistic processes similar to those found in face recognition.

Eigenfaces in Face Space

Building a machine's capability to understanding objects created from atomic representation requires machine methods that are not feature analytic. One such method is contained in the system by Turk and Pentlan [2] at the Vision and Modeling Group part of the Media Laboratory at the Massachusetts Institute of technology. This method is used to identify faces from pixel encoded pictures. The method was developed as an extension of earlier work by Sirovich and Kirby [3] to efficiently represent pictures of faces using principle-components analysis (PCA). Turk's and Pentland's system creates a face space based on eigenvectors to decompose, store, and recognize face images. The technique is a straight forward use of PCA except for an early step of image compression where the best subspace is built based on coordinates around a small set of exemplar images, termed eigenpictures. The image compression greatly reduces the calculation used in the principle components analysis. Without image compression, the analysis would need to calculate a covariance matrix the size of the number of pixels per picture squared. For instance, the number of pixel (N) in a space 256 x 256 pixels image would require calculating a matrix of the size $(65{,}536)^2$. Determining the eigenvectors and eigenvalues on such a matrix is an intractable task.

The process for face recognition using low dimension space involves the following steps:

- The first task is to select a training set (M) of pictures and create the vectors $I_1 ... I_m$ by concating the pixel intensity values. Make sure that the members of the set are normalized. Normalization is required since the analysis will be sensitive to any differences in the data set. For example, if half the pictures have dark background and half have light background then one would expect background light to be a salient dimension in the resulting analysis. Given that background light is extraneous to the task of face recognition, one would want to keep the background light the same for all pictures. A large portion of Turk and Pentland [1, 2] face recognition systems requires front-end processing of any picture to remove the background, centering of the face in the picture, assuring the faces are of the same size, etc.
- The average values are then removed from the training images by subtracting the average pixel values. Then the average face image (Ψ) is calculated by the formula,

$$\Psi = \frac{1}{M}\sum_{n=1}^{M} I_n .$$

- And, one centers the coordinate system of the training set pictures (Φ) around the average picture (Ψ) by calculating,

$$\Phi_i = \mathrm{I}_i - \Psi,$$

for i = 1 to M.

- Next one finds principal components by finding the M eigenvectors (u_n) of the covariance matrix C ,

$$C = \frac{1}{M}\sum_{n=1}^{M} \Phi_n \Phi_n^T,$$

which best describe the data.

- The kth vector is chosen such that the scalar λ_k (known as the eigenvalue) is a maximum, subject to $u_l^T u_k = 0$, for $l < k$. The formula for the eigenvalue is

$$\lambda_k = \frac{1}{M}\sum_{n=1}^{M} \left(u_k^T \Phi_n\right)^2 .$$

- The size of the matrix C is N^2. Presently, one finds difficulty in performing principle component analyses of data sets greater than one hundred. Because most pictures

have a data size or pixel number greater than one hundred, one must use a technique for principle component analysis developed by Sirovich and Kirby [3]. The analysis of Sirovich and Kirby [3] uses the fact that if the number of eigenpictures (M) used is less than the number of data points (N), and since we subtracted the mean for the data, there are only the M-1 degrees of freedom for M-1 non-zero eigenvectors. The analysis uses the M eigenpictures to create a M dimensional subspace of the possible images. Encoding into the M dimensional subspace reduces the size of the principle component analysis from N to M, and specifically the covariance matrix to the size M^2 down from N^2. Since M (the number of pictures in the training set is usually around 8 to 40) is usually much smaller than N (usually in the tens of thousands), the analysis becomes efficient. The analysis can create up to M-1 eigenvectors of which one may select M' significant eigenvectors that are sufficient for recognition. So, instead of calculating over the matrix C, one creates the M by M matrix L, Where

$$L_{mn} = \Phi_m^T \Phi_n.$$

- Finally, one finds the M eigenvectors, v_l of L, and the scalars, E_l, such that

$$E_l = \sum_{k=1}^{M} v_{lk} \Phi_k, \; l = 1,\ldots, M.$$

Recognition is accomplished by first transposing a new face into its eigenface components weights (ω) by calculating,

$$\omega_k = E_k^T (I' - \Psi),$$

for k = 1, 1,..., M'. The pattern of weights (Ω),

$$\Omega^T = [\omega_1 \ldots \omega_{M'}],$$

can be considered to be analogous to a Fourier transform of the spatial frequencies contributing to the image. Recognition is accomplished by comparing a newly transformed image to the distance,

$$d_k = \left\| \Omega - \Omega_k \right\|^2,$$

between the image and the exemplars, or other stored patterns. The new pattern is categorized as being similar to the closest stored pattern.

The eigenspace technique has been successful with images but its application to C source code is not immediately apparent. Another approach to categorizing objects by the frequency of its constituent atomic parts is the *n*-gram techniques. Important for our purposes is that the technique works directly on natural language text. Although the technique uses only the letters of the alphabet and the space symbol it can be directly generalized to C source code when one adds the additional symbols used in the language C. Next, we will examine the N-gram approach.

N-grams

N-grams are n-character sequences of text. If *n* equals one then the sequences would equal the single letters in the text, for $n = 2$ the bigrams, 3 the trigrams, etc.. The frequency of the sequences are usually collected by either sliding a "window" of the size of the *n*-gram across the text one character at a time. For example, the word "example" would generate the sequences "exa", "xam", "amp", "mpl", "ple", "le_", etc. All are sequences in a $n = 3$ *n*-gram. Using *n*-grams as a scoring technique allows modeling the statistical nature of a text in a manner that is robust to typographical errors, garbling, etc.. Some of the methods of scoring score only the *n*-gram in a word whereas others score across inter-word boundaries. At small *n*'s, the *n*-grams collect mainly information about spelling and word presence, large n's collect information on common word sequences and thus start represent the grammar of languages. Indeed, the technique using large *n*-grams can be used to generate random test in the style of the author on which the *n*-grams are collected. The approach is language independent and can be used with Japanese and Chinese computer text based on a 16-bit character code to designate symbols.

The approach was developed by Damashek [4] who used a *n*-gram system to correct spelling and typing mistakes in text. Since then, a complete system for browsing documents has been developed by Pearse and Nicholas [5] with improvements suggested by Crowder and Nicholas [6] to reduce the amount of information required for a usable system. Systems differ in implementation but have the following general features. First, the system collects the *n*-gram frequencies using a sliding "window" for a text. A *n* equal to 5 has been found to be adequate for browsing and hypertexting documents [5]. The frequencies are then normalized by dividing the n-gram counts by the total number of *n*-grams collected. Then, the counts are centered to an average text by subtracting the normalized n-gram frequencies from a corpus of average texts. The obtained vector is the document histogram that may contain thousands of entries. At this juncture one can use the data in several ways.

The Pearse and Nicholas system TELLTALE [5] determines the similarity of a text to another text with the formula

$$SIM_c(d_i, d_j) = \frac{\sum_{k=1}^{t}(d_{i,k} \cdot d_{j,k})}{\sqrt{\sum_{k=1}^{t} d_{i,k}^2}\sqrt{\sum_{k=1}^{t} d_{j,k}^2}},$$

which can be used on the normalized document *n*-gram vectors d_i and d_j to calculate the strength of the relationship between the two representations. Also one can index text by a query based on a similarity score calculated on whether document f_i or query q contain *n*-gram k by the formula,

$$SIM_i(q, f_i) = \frac{\sum_{k=1}^{t}(q_k \cdot f_{i,k})}{\sum_{k=1}^{t} q_k},$$

They also offer a method for disambiguating queries by specifying the context for a query. The context emerges from the intersection between two similarity scores: the similarity of the query and the document set, and the similarity of the current document and the

document set. The disambiguation of the set (d) requires the ad hoc setting of thresholds for the two similarity judgments.

$$SET_d = SET_c \cap SET_l \text{ Given that}$$

$$SET_l = \{d_i | SIM_l(q, d_i) > threshold_l\} \text{ and } SET_c = \{d_i | SIM_l(q, d_i) > threshold_c\}.$$

Eigencodes in Codespace

Both the n-gram and the Eigenspace approaches seek to identify groups of objects. Both approaches seek to capture the frequency of the atomic features in a representation. The difference between the two come from the analytical techniques used to determine grouping. To date, eigenfaces are formed on dimensions extracted from the data using principle components analysis with grouping determined by vectorial differences in the space, and n-grams determine grouping or proximity primarily through cluster analysis or proximity based on least squared differences of features. Again, cluster analysis methods are interested in identifying groups of objects in whatever dimension space the objects are in, whereas factor analytical methods are interested primarily in the dimension of the space the objects are in. Principle components analysis has the advantage of reducing a large number of variables to a smaller number of variables (locations of axis) for further analysis. The reduction, if done appropriately, reduces noise, gives storage economy, and allows the identification of underlying variables or those dimensions on which the C source code varies.

The approach chosen by the author to create a categorization technique for C source code is to use the low dimensional eigenface technique, over the frequency counts of the n-gram approach. This should be achievable since both techniques uses counts or intensities of atomic representations in a vectored format. Given that the data are comparable, one may apply the eigenface technique to the n-gram counts of C source code. In other words, to look for eigencodes in a code space and analyze the reasonableness of the space

for representing code. The general approach is to make the codespace require the following steps:

- Acquire a small representative set of C source code to form the exemplar set on which to build eigenspace for the code.
- Count the N-gram frequencies for each source of code, and place them in a vector.
- Normalize the frequencies by dividing by the total N for each vector.
- Center the vector around the average vector.
- Apply large data set conversion if necessary.
- Do principle components analysis to find the eigenvalues and eigenvectors to describe the lower dimension subspace of the eigenspace.

METHODOLOGY

To test the concept of building a cognitive filter for C source code, the author selected an approach where the representation of the code would be captured in n-grams. The encoding of source code into *n*-grams allows the code to be expressed in atomic representations. The representation is the set of symbols that are used to form C code and the conditional probabilities of a code being selected given the proceeding *n*-1 symbols. Such a representation is suitable to represent any C source code as well as any general text. The recognition process over the *n*-grams used the eigenspace low dimensional categorization technique used in the face recognition. This was chosen over the clustering techniques because the eigenspace systems allow a coherent approach to categorization in which the underlining space can be understood in terms of its structure. In particular, one might want to look for a piece of C source code for which one does not have a good example, but can describe by its probable location in the eigenspace dimensions.

A group of eight software programs were selected to test the ability of a filter to derive dimensions for an eigenspace that would result in meaningful categorization of C code. The programs were selected based on the criteria that the programs performed a single

function and the function of the code was either system utility, mathematical, statistical, or logical. The users comments and extraneous lines were removed from the code. The programs were then encoded in *n*-grams with an *n* of one (i.e., symbol frequency). All possible one-gram symbols were first collected and assigned a position in a vector. Each vector was normalized by calculating the *n*-gram likelihood for each program by dividing each *n*-gram by the total number of *n*-grams in the program. The average *n*-gram vector was calculated. A principle components analysis was performed given 7 factors or dimensions for the code space. A characterization of the dimension was achieved by the author by examining the scorings weights for the symbols.

The results are surprising in that the first and fourth dimensions, which accounted for 40 percent of the trace variance, were related to the text and programming style of the programmer. The first factor, which accounted for the greatest variance in the space, was whether the programmer used indentation as a textual cue for nesting lines in the program or writing with all lines left justified. The fourth factor was whether the programmer structured the code in functions and defines as compared to straight coding. The other factors were tied to the content of the programs. Factors two and three was related to input-output or internal processing. Factor two was a dimension of input/output software versus internal processing, and factor three was whether input-output processing was string or character. Factor 5 was code using processes based on logical comparison or not. Factor 6 was numerical processing or not. Factor seven was whether systems software used macro defines or not.

The findings seemed to be at odds with the purpose of the cognitive filter, particularly, that the programmer's use of indentation was a major factor. So, a second attempt at creating the eigenspace was tried with the same procedure but with the removal of all preceding spaces for a line of code. The assumption was that by removing the spacing the

analysis would center around other variables. The result was that structured versus unstructured programming style became intertwined with the other factors making understanding of the dimension difficult. The first two factors where anchored on one end by the most structured and nested of the programs, and on the other end with either an unstructured math program (factor one) or unstructured logic program (factor two). The first analysis seemed to create a much more coherent space. And, the style of programming does seem to make a marked difference in the frequency the *n*-grams.

Summary

The code space approach to building a low dimensional filter is a promising approach to categorizing software source code. The ability to produce a coherent, understandable space in which to search will be advantageous to automated software reclaiming techniques and is essential for human use. The surprising finding that the programmer's style is a major dimension detracts from the coherency of the space. However, that finding is not all bad. The knowledge that a piece of software is programmed in a structured format may well be useful knowledge for any search or post filter analysis of the software. For example, formal methods that decompose legacy software to understand its functions or intentions might work best with structured code. Humans might want to limit searches to structured code so that any software found will be easier to comprehend and to verify as useful.

Further development of a cognitive filter to aid in the search should focus on several aspects. The filter should be extended to include a parsing function to allow greater flexibility in normalizing the code for removing extraneous differences. For instance, one might equate types of iteration using goto's or for's, regardless of how it was programmed. The addition of *n*-gram techniques, particularly key word look up, would be important for giving the filter the ability to make fine discriminations in the software. A programmer might be interested in all of the programs that use Boltzmann's constant, or a programmer

might be interested in redefining certain system files. The technique could allow users to select their own examples to create similarity sets upon which to conduct searches and build descriptions.

The addition of the n-gram techniques could provide a means of using the programmer's comments. The eigencode technique strips the programmer's comments since they act as a source of unwanted variation to add to the written code. But by using the n-gram techniques on the user comments, one could add the ability to use the programmer's description of their goals and intentions to help find software. The problem with using the n-gram techniques is that most of the n-gram techniques are computationally expensive, since they require on-the-fly comparisons of many members, sometimes all members, of a stored database. The eigenspace technique's strongest advantage may well be as a prefilter to limit and order the search process, followed by a focused use of the more computational n-gram technique in the most fruitful areas.

REFERENCES

[1] M. Turk and A. Pentland, "Face processing: Models for recognition, "*Intelligent Robots and Computer Vision VIII*, Proc. SPIE Vol. 1192, (1989), pp. 22-32.

[2] M. Turk and A. Pentland, "Recognition in face space, *"Intelligent Robots and Computer Vision IX: Algorithms and Techniques"*, Proc. SPIE Vol. 1381, (1990), pp. 43-54.

[3] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Am. A*, Vol. 4, No. 3, Mar. (1987), pp. 519-524.

[4] M. Damashek, " Gauging similarity with *n*-grams: Language-independent categorization of text," *Science*, Vol. 267, Feb. (1995), pp. 843-848.

[5] C. Pearce and C. Nicholas, "TELLTALE: Experiments in a dynamic hypertext environment for degraded and multilingual data," *J. Am. Soc. Inf. Sci.*, Apr. (1996).

[6] G. Crowder and C. Nicholas, "Using statistical properties of text to create metadata," *Proc. of the First IEEE Metadata Conference*, Apr. (1996).

[7] R. Noel and S. Acchione-Noel, "Holistic processes in evolution: Evolving non-decomposable images using evolutionary computation," (manuscript in review).

[8] C. Green, D. Luckham, R. Balzer, T. Cheatham, and C. Rich, "Report on a knowledge-based software assistant," *Final Technical Report RL-TR-83-195*, Rome Laboratory, Aug. (1983)

[9] M. Gerken, N. Roberts, and D. White, "The knowledge-based software assistant: A formal, object oriented software development environment," Proc. IEEE Nat. Aero. & Elec. C., May (1996), pp. 511-518.

[10] M. Chase, D. Harris, S. Roberts, and A. Yeh, "Analysis and presentation of recovered software architectures," *KBSE 96*, In Press (1996).

[11] V. Bruce and G. W. Humphreys: "Recognizing objects and faces." *Visual Cognition 1 (2/3)*: 1994, pp. 141-180.

[12] D. Marr and H. K. Nishihara: "Representation and recognition of the spatial organization of three-dimensional shapes." *Proceedings of the Royal Society of London B200*: 1978, pp. 269-294.

[13] I. Biederman: "Recognition-by-components: A theory of human image understanding." *Psychological Review 94*: 1987, pp. 115-147.

[14] S. Ullman: "Aligning pictorial descriptions: An approach to object recognition." *Cognition 32*: 1989, pp. 193-254.

[15] C. J. Price and G. W. Humphreys: "The effects of surface detail on object categorization and naming." *Quarterly Journal of Experimental Psychology 41A*: 1989, pp. 797-828.

FREQUENCY RESPONSE OF SEMICONDUCTOR PHOTOREFRACTIVE MATERIALS: ZnTe:Mn:V, GaAs:Cr, and CdMnTe:V

Jeffrey B. Norman
Associate Professor
Department of Physics and Astronomy

Vassar College
124 Raymond Avenue
Box 559
Poughkeepsie, NY 12601

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC

and

Rome Laboratory

August 1996

FREQUENCY RESPONSE OF SEMICONDUCTOR PHOTOREFRACTIVE MATERIALS: ZnTe:Mn:V, GaAs:Cr, and CdMnTe:V

Jeffrey B. Norman
Associate Professor
Department of Physics and Astronomy
Vassar College

Abstract

The photorefractive frequency response of three compound semiconductor materials, ZnTe:Mn:V, GaAs:Cr, and CdMnTe, were studied, using two beam coupling with a moving intensity grating, and no externally applied electric field. The data exhibit significant deviations from the expected Lorentzian response based on the single trap, single charge species photorefractive model. A sharply peaked behavior at low frequency is observed in all three cases, suggesting that its source is related to material properties that are particularly common to semiconductor photorefractives. Although the mechanism behind this unexpected frequency response has not yet been explained, possible models are presented and experimentally tested here and an effort to explain our results through numerical solutions of the nonlinear photorefractive equations is continuing.

FREQUENCY RESPONSE OF SEMICONDUCTOR PHOTOREFRACTIVE MATERIALS: ZnTe:Mn:V, GaAs:Cr, and CdMnTe:V

Jeffrey B. Norman

Introduction

Photorefractive semiconductors comprise an important class of photorefractive materials because of their potential for practical applications in optically-based signal and image processing systems. Two of their most important properties in this regard are sensitivity in the 0.6-1.3 μm near-infrared region (and their resulting compatibility with diode lasers) and their sub-millisecond response times using moderate intensity cw lasers. This class of materials includes appropriately doped CdTe, GaAs, InP, GaP, CdS, ZnTe, and CdMnTe. ZnTe and CdMnTe are the most recently grown and studied of these materials [1,2].

In order to optimize the properties of photorefractive materials for applications, it is necessary to accurately measure their microscopic properties relevant to photorefractive behavior, such as trap concentrations and electro-optic coefficients. Until such properties are combined with a material's wavelength, spatial frequency, and temporal frequency responses, optimization of material growing conditions and realistic evaluations of application potential cannot be made.

Samples of undoped ZnTe, ZnTe:V, and ZnTe:Mn:V have all exhibited photorefractivity [2]. Of these, ZnTe:Mn:V has the largest measured two-beam coupling gain. The initial purpose of this project was to perform two-beam coupling measurements on a particular sample of ZnTe:Mn:V for the determination of essential photorefractive properties. However, in the course of the experiments, anomalous behavior was discovered in its frequency response, and subsequently similar behavior was found in crystals of GaAs:Cr and CdMnTe:V. This led us to expand the scope of the project to a study of the mechanism behind these observations. Therefore, in addition to a detailed characterization of the ZnTe:Mn:V crystal, we present here an effort to experimentally characterize and model the anomalous time response observed in all three crystals.

Methodology

Since the photorefractive time constants of the semiconductor samples is in the sub-millisecond regime, at moderate light intensities, it is convenient to measure this response in the frequency domain rather than in the time domain, thereby alleviating the need for fast shutters and photodetectors. The frequency domain method is entirely in the steady state regime. It is normally accomplished by subjecting the crystal to a linear intensity grating moving with constant velocity perpendicular to the fringes. By measuring the steady state gain as a function of the grating velocity, the frequency response is obtained, and from the predicted Lorentzian dependence the photorefractive time constant is normally extracted. All of the data obtained in this project were taken at a wavelength of 1.064 μm and some of the data were compared with earlier results at 0.75 μm.

Two experimental configurations for nonstationary two beam coupling have been used in this work. In the first, shown in Fig. 1, the moving grating was produced by a linear phase modulation of one of the interfering beams using an electro-optic phase modulator. The steady-state two-beam coupling gain was measured as a function of the frequency, f, of the ramp applied to the e-o modulator. The velocity of the grating is simply related to this frequency by $v_g = \Lambda \cdot f$, where Λ is the grating spatial period, since the ramp amplitude was adjusted for a 2π phase excursion. Removal of mirror M3 allowed monitoring of the fringes using the resulting Mach-Zehnder configuration.

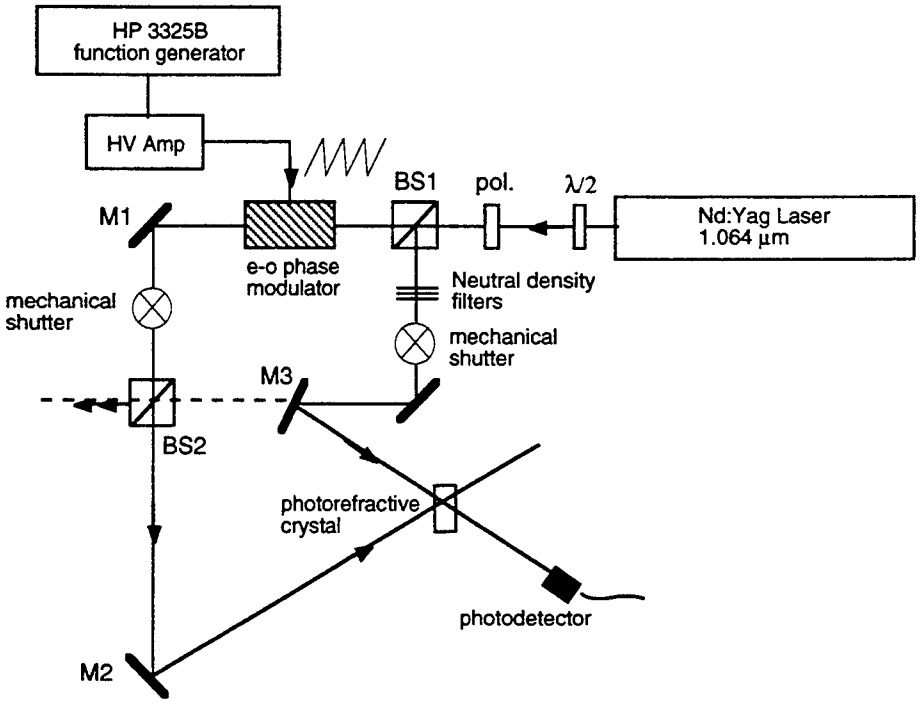


Figure 1. Electro-optic configuration for two-beam coupling measurements.

The second configuration, shown in Fig. 2, was used as an independent check on the results obtained from the first, to eliminate the possibility of any system-generated effects. In this method, acousto-optic modulators driven by phase-locked signal generators were used to produce a relative frequency shift between the beams. The grating velocity is then proportional to the frequency detuning, δf, between the two drive signals by $v_g = \Lambda \cdot \delta f$. It was found that the results from the electro-optic and acousto-optic approaches were consistent with each other.

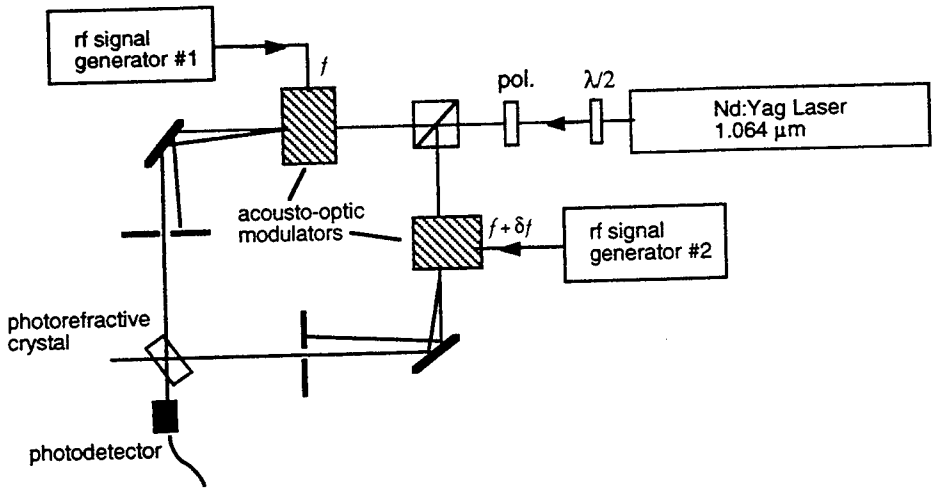


Figure 2. Acousto-optic configuration for two-beam coupling measurements.

In each case, the laser beams were incident on the $(11\bar{0})$ face of the crystal, the grating vector was oriented in the $\langle 001 \rangle$ direction, and the beams were s-polarized along the $\langle 110 \rangle$ direction in order to take advantage of the r_{41} electro-optic coefficients. The dimensions of the three crystals, in the directions $\langle 110 \rangle \times \langle 100 \rangle \times \langle 11\bar{0} \rangle$, were: GaAs:Cr - 11x10x6 mm^3, ZnTe:Mn:V - 4x3x1.5 mm^3, CdMnTe:V -

Experimental Results and Modeling

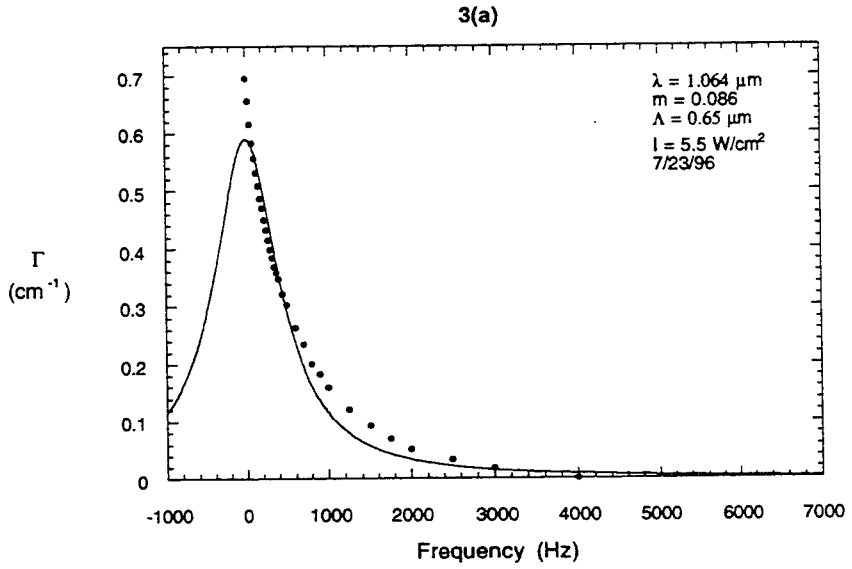
Representative frequency response data for the ZnTe:Mn:V sample is shown in Fig. 3. In Fig. 3(a), the data are fit to a single Lorentzian function,

$$\Gamma = \frac{\Gamma_0}{1 + (2\pi f \tau)^2} \quad , \qquad \text{Eqn. (1)}$$

where Γ is the exponential gain coefficient, τ is the photorefractive time constant, and f is the frequency applied to the electro-optic modulator (or, alternatively, the frequency detuning between the acoust-optic modulators) . Eqn. (1) is the form predicted by the single trap, single charge species model. It is immediately clear that, although the external experimental conditions lead us to predict a single Lorentzian response, the data deviate significantly from this picture. Since the transient response of photorefractive materials is crucial to a number of applications in signal

and image processing, it is important that it be well characterized and modeled. We have therefore considered a number of models to explain this behavior. These models and our experimental and theoretical tests of them are described in what follows.

The possibility of electron-hole competition explaining these results appears to be readily eliminated, since this is expected to decrease rather than increase the gain at low frequencies and would also lead to characteristic effects in the spatial frequency response which we do not observe (see Fig. 7). Another model which we have hypothesized is one in which the crystal possesses two distinct sets of photorefractive trap levels, displaying widely different temporal behavior from each other. If one assumes, as a first approximation, that the gratings associated with these two sets of levels respond independently of each other, each in a Lorentzian manner, it follows that the observed frequency response might be made up of the sum of two Lorentzians. Although this picture neglects interactions between the levels, which is certainly occurring through the free charge carriers, it may approximately describe the behavior. Fig. 3(b) shows the fit of the data to a sum of two Lorentzian functions. The quality of the fit in Fig. 3(b) leads us to believe that this model may have some merit, or at least that it cannot be yet be eliminated . Δ_1 and Δ_2 are the full widths at half maxima of the two Lorentzian components, as determined from the fits and τ_1 and τ_2 are the corresponding time constants.



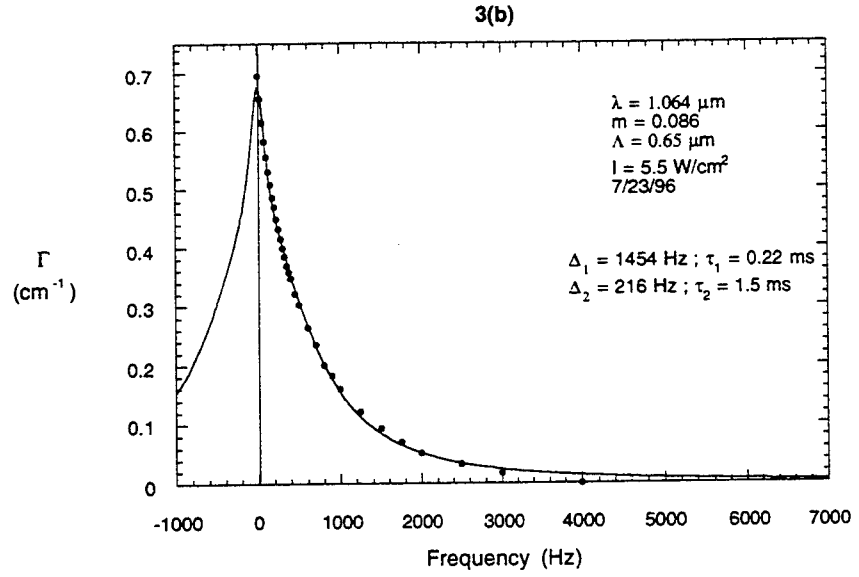


Figure 3. Frequency response data for ZnTe:Mn:V. (a) Single Lorentzian fit, (b) Double Lorentzian fit.

Another model considered was one in which there is an anomalous frequency-dependent variation in the photorefractive phase shift, that is, the phase shift between the intensity and photorefractive gratings. For maximum two beam coupling gain, this phase shift must be $\pi/2$. Even the predicted Lorentzian response takes into account a frequency-dependent phase shift, ϕ,which is equal to $\pi/2$ at 0 Hz and which decreases at nonzero frequencies as $\phi = \pi/2 - \tan^{-1}(2\pi f\tau)$, where τ is the photorefractive time constant. The possibility was considered that a departure from this functionality might account for the anomalous frequency response behavior. As shown below, this explanation is unlikely in light of subsequent data.

The photorefractive frequency response of GaAs:Cr and CdMnTe:V, have been measured and compared with ZnTe:Mn:V. These data are shown in Figures 4 and 5. The GaAs:Cr data consist of manually obtained discrete points, while the CdMnTe:V data were obtained by continuously sweeping the function generator frequency in the configuration of Fig. 1.

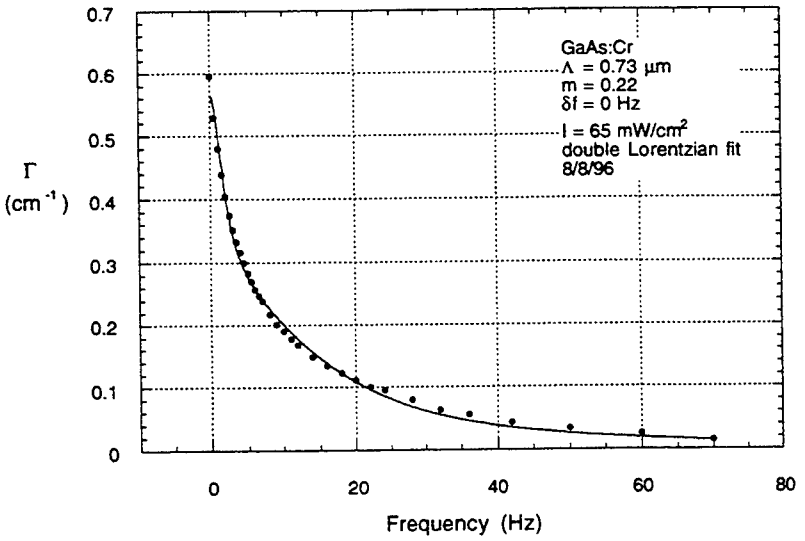


Figure 4. Frequency response of the GaAs:Cr sample.

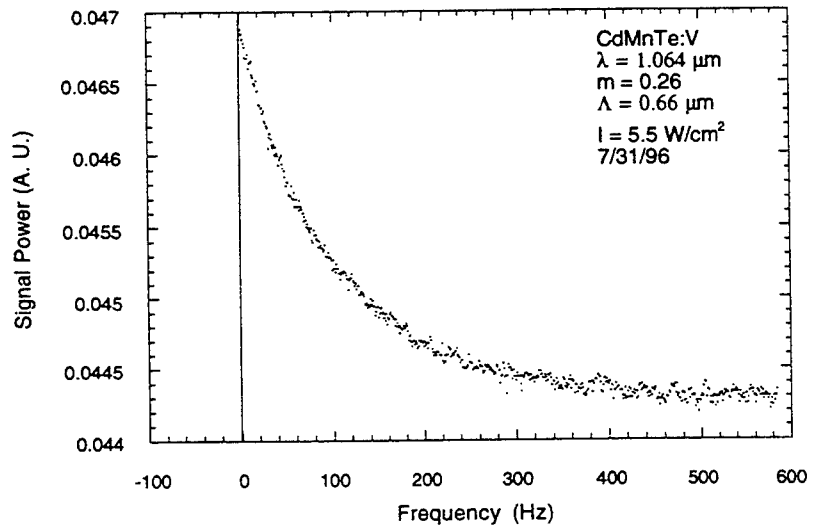


Figure 5. Frequency response of the CdMnTe sample.

It is clear from the data displayed in Figures 4 and 5 that the GaAs:Cr and CdMnTe:V samples exhibit similar frequency response behavior to that of ZnTe:Mn:V. This leads us to the question of whether the explanation for this effect might be found by

considering material properties that are particularly common in semiconductor photorefractives.

We have also measured the temporal behavior of these materials in the time domain, that is, by subjecting them to a stepped intensity input. This is just the time domain version of the frequency response data. The GaAs sample was chosen as the primary material for the time domain study because of its large effective gain. These measurements require that the time constant of the measurement system be significantly faster than the rise time of the photorefractive grating. For this reason, the intensity was kept very low in the time domain measurements. Because of this requirement and that of low grating modulation depth, the signal from the ZnTe:Mn:V crystal was plagued by noise. The larger effective gain of the GaAs:Cr sample (because of its larger thickness) increased the signal to noise ratio to an acceptable level. As shown above, all three crystals have shown the same frequency response behavior, so time response data on any one of them was deemed equally valuable.

The time response of the GaAs:Cr sample is shown in Fig. 6, where degenerate beams were used (i.e. stationary grating), while the remaining external conditions were identical to those of Fig. 4. The dashed line corresponds to the best fit of the data to the transient behavior, assuming a single time constant. Significant departure from this model is evident. Similar behavior was seen with the ZnTe:Mn:V sample, though the data show considerably more noise. The solid line represents a double exponential fit, which one would expect to be valid in the presence of two independent photorefractive time constants. The fit to the data is good and is consistent with the similar quality of the fit of the frequency response data to a double Lorentzian function in Fig. 4. These data lend credibility to the multiple trap species model and lead us to discount the possibility of an anomalous frequency dependent photorefractive phase shift.

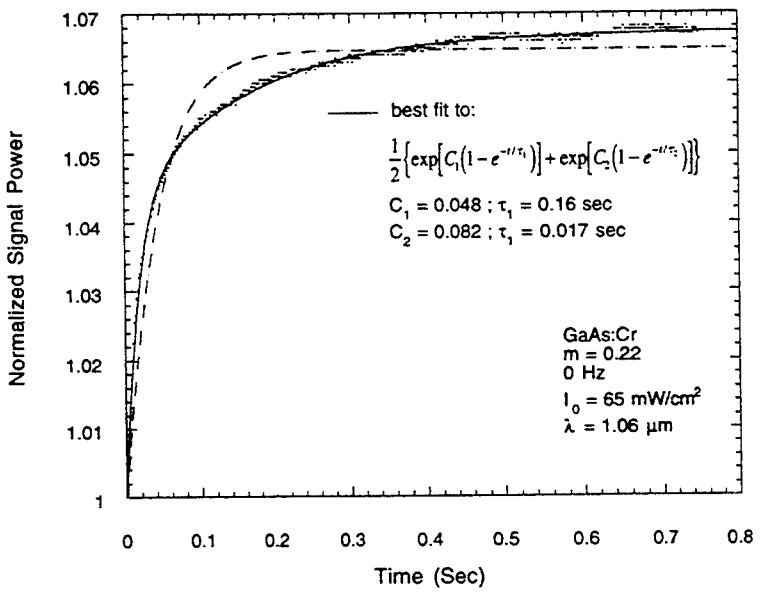


Figure 6. Time response of the GaAs:Cr sample. Dashed line: single exponential growth, solid line: double exponential growth.

Other frequency response data have been obtained in all three samples, including a wide range of grating periods, total light intensities, and grating modulation depths and in every case the sharply peaked behavior was observed and the data fit relatively well to the sum of two Lorentzian functions.

The spatial frequency response of the ZnTe:Mn:V sample at 1.064 µm has been measured and is shown in Fig. 7, along with similar data taken earlier at 0.75 µm. The expected inverse dependence of the gain on wavelength was approximately observed and, somewhat surprisingly, no deviation from the single trap, single charge species model of the grating period behavior, shown by the solid line fits, can be detected in these data. The inverse wavelength dependence of the gain is an indication that the dominant charge carrier does not change sign. Note the close agreement obtained for the electro-optic coefficient of the material, r_{eff}, at the two wavelengths, and that the values obtained for the trap concentrations, N_E, agree within a factor of two. We do not think that this difference in concentrations is significant, in light of the sensitivity of the fit to this parameter. Significantly,

however, the values obtained for the electro-optic coefficient both differ greatly from the accepted bulk value of about 4.2 pm/V. We have no explanation for this discrepancy, but it leads us to the question of whether or not it is related to the anomalous temporal response. A direct bulk measurement of the electro-optic coefficient for this particular crystal has not been made.

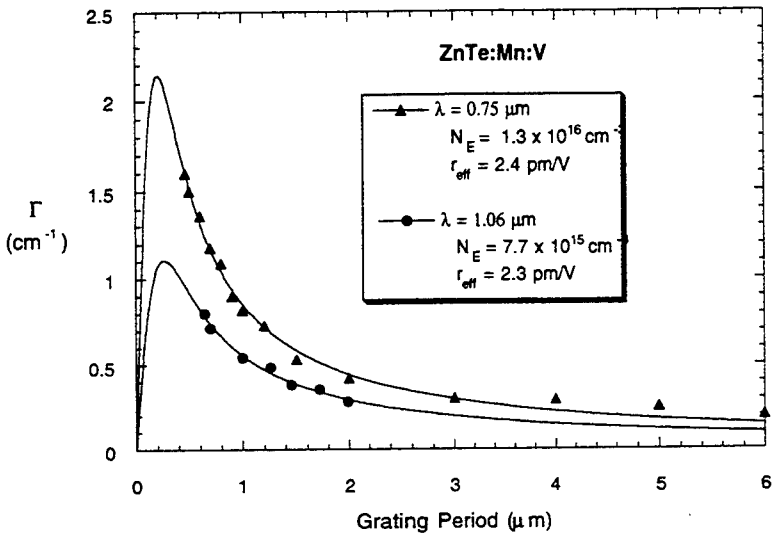


Figure 7. Spatial frequency response of our ZnTe:Mn:V sample.

We have also performed a systematic study of the temporal response of our ZnTe:Mn:V sample as a function of grating period. Our procedure was to fit the frequency response data obtained at each grating period to a double Lorentzian function and then to infer the time constants of the two hypothesized photorefractive components, which are inversely proportional to the widths of the two Lorentzians, from the constants of the fit. Under the assumptions of the single trap, single charge species model, one expects the photorefractive response rate to increase with increasing spatial period. This is not what is observed in our ZnTe:Mn:V sample, as seen in Fig. 8. In fact, there seems to be a pronounced minimum in the time constants for both the fast and slow components, although at different grating periods in the two cases.

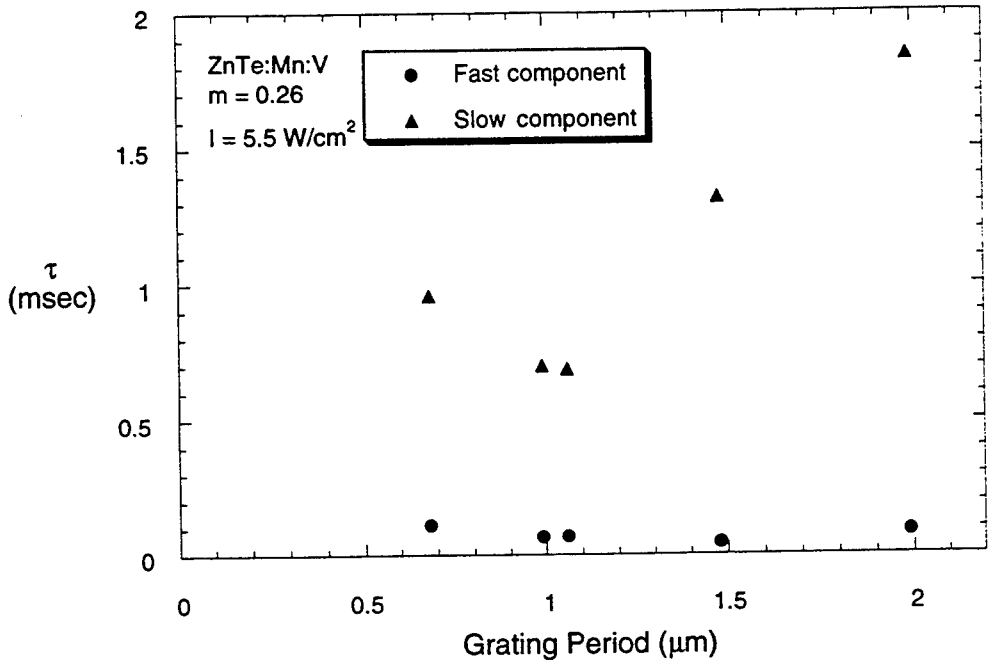


Figure 8. Spatial period dependence of the slow and fast components of the photorefractive grating in ZnTe:Mn:V.

As an additional clue to the origin of our temporal response data, measurements were made of the spatial period dependence of the gain at various frequencies, which are shown in Fig. 9. These data are an extension of those in Fig. 7, which displays only 0 Hz data.

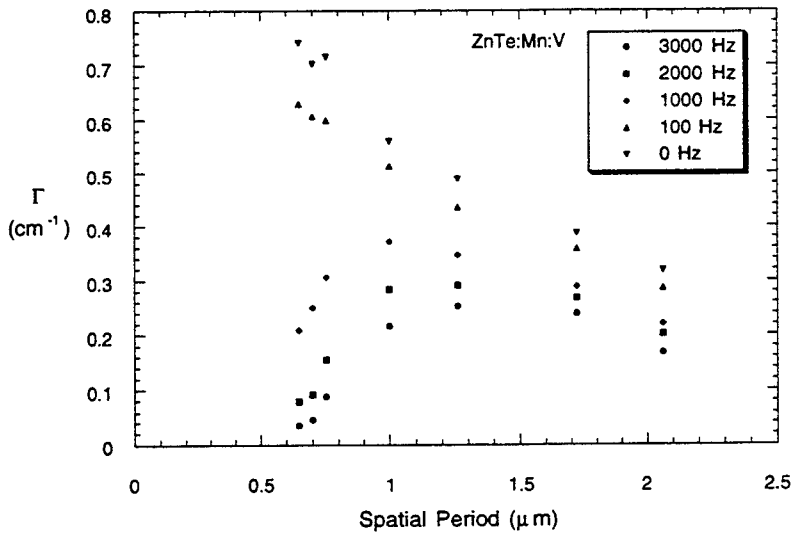


Figure 9. Spatial period response of ZnTe:Mn:V at various frequencies. The solid lines are meant only to guide the eye.

The spatial frequency response of a nonstationary grating is known, under the normally assumed approximate conditions. The spatial frequency response can be written

$$\Gamma \propto \frac{K\left(1+K^2/k_0^2\right)}{\left(2\pi f\tau_d\right)^2\left(1+K^2/\bar{k}^2\right)^2+\left(1+K^2/k_0^2\right)^2} \quad . \qquad \text{Eqn. (2)}$$

The data of Fig. 9, however, are not well described by this functional form.

Discussion

Based on the experimental results presented above, the most promising model for explaining the observed anomalous temporal response of our semiconductor photorefractive samples seems to be one which includes multiple trap species. This is rather unexpected given the wide range of species which one would expect to be present in the three materials ZnTe:Mn:V, GaAs:Cr, and CdMnTe:V. Nevertheless, both the frequency and time response data indicate multiple time constant behavior

and having shown that electron-hole competition and anomalous frequency-dependent photorefractive phase shifts are unlikely candidates, the multiple trap level species model is the one that appears to be most likely at this time. This is in spite of the fact that Ziari, et al. [1] found a single exponential time response to satisfactorily describe the behavior of their ZnTe:V sample. We are in the process of numerically modelling the temporal behavior of these crystals using a two trap species model.

Acknowledgments

I am grateful to the Summer Faculty Research Program of the Air Force Office of Scientific Research and to Rome Laboratory for making this work possible and to Andy Pirich and Dr. George Brost of Rome Laboratory for facilitating and collaborating on this project.

References

1. M. Ziari, W. H. Steier, P. M. Ramon, S. Trivedi, and M. B. Klein, "Photorefractivity in vanadium-doped ZnTe," Appl. Phys. Lett. 60, 1052 (1992).

2. S. B. Trivedi, J. I. Soos, Y. K. Sun, R. D. Rosemeier, W. H. Steier, M. Ziari, R. N. Schwartz, M. B. Klein, "Vanadium doped ZnTe, a new material for real-time optical signal processing: crystal growth and optical characterization," United States Air Force, Phase I Final report, contract F30602-94-C-0115, 1994.

3. P. Yeh, *Introduction to Photorefractive Nonlinear Optics* (Wiley, New York, 1993).

RAPID PROTOTYPING OF SOFTWARE RADIO SYSTEMS USING FIELD PROGRAMMABLE GATE ARRAYS AND DSP MICROPROCESSORS

Glenn E. Prescott
Associate Professor of Electrical Engineering
Department of Electrical Engineering & Computer Science

University of Kansas
1013 Learned Hall
Lawrence, KS 66045

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling AFB, DC

and

Rome Laboratory

August 1996

RAPID PROTOTYPING OF SOFTWARE RADIO SYSTEMS USING FIELD PROGRAMMABLE GATE ARRAYS AND DSP MICROPROCESSORS

Glenn E. Prescott
Associate Professor
Department of Electrical Engineering & Computer Science
University of Kansas

Abstract

Field programmable gate arrays (FPGA) and monolithic DSP microprocessors are powerful technologies which can be used to maximum advantage in military software radio applications. The objective of this report is to examine the role of both of these technologies in the implementation of high performance military radio systems. A radio transceiver implemented using state-of-the-art DSP technology - often referred to as *software radio* - requires real time signal processing at a variety of bandwidths. In order to accommodate the needed bandwidths in a discrete time implementation, it is appropriate to use devices which are well suited to each stage of the system - fast, yet algorithmically simple devices for the wide bandwidth stages and slower, yet more flexible devices for the processing required at low bandwidths. This report examines the processing requirements of software radio, and assesses the role of the current generation FPGA, and DSP processor technology in implementing the algorithms required to make these radio systems function efficiently.

RAPID PROTOTYPING OF SOFTWARE RADIO SYSTEMS USING FIELD PROGRAMMABLE GATE ARRAYS AND DSP MICROPROCESSORS

Glenn E. Prescott
Associate Professor of Electrical Engineering
Department of Electrical Engineering & Computer Science
University of Kansas, Lawrence, KS 66045

1. INTRODUCTION

Until recently radio transmitters and receivers were almost exclusively implemented with analog electronic components. However, a new approach is now becoming popular - one that employs digital electronics to implement most of the analog signal processing functions in the radio. This evolution in radio system design is driven by the ever increasing speed and decreasing cost of microprocessors and high performance analog-to-digital (ADC) and digital-to-analog (DAC) converters. It is no longer uncommon to sample a received signal at the intermediate frequency (IF) stage and process the signal with numerical algorithms using specialized digital signal processing (DSP) hardware. The DSP hardware performs a variety of operations on the signal including down conversion, demodulation, and filtering; all of which are inherently continuous-time (i.e., analog) processes.

1.1 Digital Signal Processing: Capabilities and Requirements

The mathematics of digital signal processing provides the framework for the design of software radio algorithms, while modern high speed digital electronic components make real time implementation of these algorithms possible. However, the hardware currently available to implement DSP algorithms for all stages of the radio system is still limited in speed, accuracy and flexibility. Initially, digital signal processing was used only for baseband waveform processing. As digital electronic devices increased in speed, DSP was soon applied to signal processing functions performed at higher frequencies - e.g., the final IF stage in a radio receiver. Functions such as IF bandpass filtering, automatic gain control (AGC), and coherent modulation and demodulation are typically required at this stage. In the absence of a sufficiently high speed processing capability, innovative techniques such as sub-sampling are used to process bandpass signals of small to moderate bandwidth. This has allowed the boundary between analog and digital processing to be pushed as far up the signal path towards the antenna as permitted by physical electronic devices. For most types of moderate data rate communications - on the order of 100 KB/s or less - bandwidth is not a serious barrier to DSP techniques. However, military

radio systems pose a notable challenge because of the wide bandwidth characteristics of spread spectrum modulation.

1.2 Military Radio Signal Processing Requirements

Military communication systems often require the use of spread spectrum techniques to provide an antijam (AJ) capability, or some measure of covertness through the use of low probability of intercept (LPI) waveforms. The result is that extremely wide bandwidth signals are present at the output stage of the transmitter and the input stages of the receiver. We know from the Nyquist theorem and fundamental bandpass sampling techniques that bandpass signals can be sampled at a rate no less than the bandwidth of the signal; so high frequencies alone do not put a limitation on DSP processor capability. However, wide bandwidth signals are a challenge for any type of digital signal processing hardware, and they are especially troublesome for conventional DSP microprocessors. While conventional DSP microprocessors are optimized for real-time data processing, they are nevertheless implemented using the traditional von Neumann architecture - an inherently serial architecture which uses a single multiplier and executes one instruction at a time. While providing the advantage of flexibility through programmability, this architecture limits the speed with which signal samples can be processed. Even modern DSP microprocessors operating at 40 million instructions per second (MIPS) have a useful bandwidth limit of less than 500 KHz. This is especially troublesome for military communication systems which employ AJ and LPI waveforms having typical bandwidths in excess of 10 MHz.

1.3 Advantages of Specialized Digital Hardware

When digital signal processing at wide bandwidths is required the radio designer turns to specialized hardware which can operate at much higher throughputs than is possible with a DSP microprocessor. These include application specific standard products (ASSP), application specific integrated circuits (ASIC), and field programmable gate arrays (FPGA):

Application Specific Standard Products (ASSP) such as FIR filters, correlators, and FFT processors, permit certain popular DSP algorithms or functions to be optimized in hardware at the cost of flexibility. Use of ASSPs can significantly increase the device count and often presents special interface problems which can lead to further complications. Furthermore, due to a narrow range of applicability, many ASSPs may not be available in state of the art process technology [1].

When performance is a factor and product volume is high, many designers turn to ASIC technology. ASIC technology offers the ability to design a custom architecture that is optimized for a particular application. For example a conventional DSP microprocessor has only a single

multiply-accumulate (MAC) stage (see Section 3), so each filter tap must be executed sequentially. An ASIC implementation of a DSP algorithm, on the other hand, might have multiple parallel multiply-accumulate (MAC) stages. When comparing the performance of the ASIC versus the DSP microprocessor it becomes apparent that the DSP microprocessor offers slow speed but maximum flexibility (due to programmability) while the ASIC provides high speed with minimal flexibility. Between these two extremes lies the field programmable gate array [2].

1.4 Field Programmable Gate Arrays

Modern field programmable gate arrays can implement functions beyond the capabilities of today's DSP microprocessors. In fact, they have the potential to provide performance increases of an order of magnitude or better over traditional DSP microprocessors, but with the same flexibility [3]. These devices can provide the programmability of software, the high speed of hardware and can be reconfigured in-circuit with no physical change to the hardware. In fact, FPGAs are really "soft" hardware, in that they are a good compromise between flexible all-software approaches which unfortunately limit throughput, and custom hardware implementations, which are more expensive and inflexible [4]. FPGAs offer a powerful approach - an architecture tailored to the specific application. Because the logic in an FPGA is flexible and amorphous, a DSP function can be mapped directly to the resources available on the device. Modern FPGAs have sufficient capacity to fit multiple MACs or algorithms into a single device along with the interface circuitry required by the application - a single chip solution.

Although FPGAs can out-perform DSP microprocessors under some circumstances, they are not universally the best choice for processing at every stage of the software radio. The limitations and advantages of FPGAs compared to those of the DSP microprocessor are examined further in the sections that follow. At the conclusion of this report, a suggestion is presented for the use of both the FPGA and the DSP microprocessor in a software radio testbed.

2. SOFTWARE RADIO

The essential concept of software radio is that most of the analog signal processing operations of the radio transmitter and receiver are implemented with digital hardware using DSP techniques. The placement of the receiver analog to digital converter (ADC) and the transmitter digital to analog converter (DAC) as close to the antenna as possible are distinguishing characteristics of the software radio. In the software radio receiver, the approach often used is to digitize an entire band and to perform IF processing, baseband, bit stream and other functions completely in software [5]. This approach requires the use of high speed analog to digital

converters and high speed DSP microprocessors. However, the signal processing requirements for military and commercial radio systems employing high data rate signals or spread spectrum modulation easily exceeds the processing speeds currently available in off-the-shelf DSP microprocessors. In this case, special purpose DSP hardware, application specific devices and field programmable gate arrays can play an important role.

The motivation for implementing radios in software is that a highly flexible and reconfigurable communication system can be implemented for relatively low cost. The ability to adapt the radio to its environment by changing filters, changing modulation schemes, switching channels, using different protocols and dynamically assigning channels and capacity are features which are impractical to deliver with hardware alone. Since the behavior of the software radio can be changed so easily, defining a particular architecture does not limit the radio to one specific function. Instead multiple radio systems can share a common front-end analog radio tuner while having independent digital processing for each individual radio channel [5].

2.1 A Software Radio Architecture

A software radio is essentially a hybrid analog and digital processing system. As illustrated in Figure 1, fixed analog filtering and frequency conversion are still used in the RF stages. Conceivably, there will always be a need for an analog low-noise preamplifier to capture the signal from the antenna and establish the noise figure for the receiver. Also, a down conversion operation which places the signal at some convenient intermediate frequency and allows for additional conditioning of the signal before sampling will probably continue to be a part of the software radio system for the next decade.

Using a sufficiently fast DSP microprocessor, a single device could be used to process the signal through all stages of the communication system. However, the signal processing requirements for each stage are quite different. In the IF stages, relatively simple high speed digital processing is needed, and special purpose DSP hardware can used to satisfy this requirement. At this stage, signal processing is usually limited to filtering, correlation or FFT processing. At the baseband stage the spread spectrum modulation has been removed and the bandwidth of the signal is much narrower, meaning that fewer samples need to be processed per unit time. However, the complexity of the algorithms required at this stage increases dramatically, and the extra time between samples is required in order to implement digital phase locked loops and other computationally intensive algorithms. Use of simple, high speed DSP processing at the wide bandwidth stages and slower, more flexible processing at the lower bandwidth stages will efficiently satisfy both the complexity and high throughput requirements of modern radio systems [4].

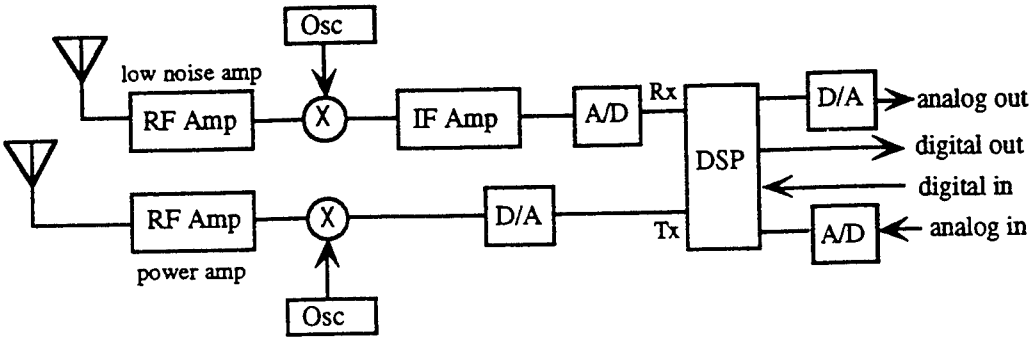


Figure 1 - Software Radio System

2.2 Software Radio Processing Requirements

The single most critical requirement in software radio is real-time processing. If the system is to operate in real time, then the data must be moved in and out of the DSP microprocessor on a regular (i.e., sample by sample) basis, where hundreds of instructions may need to be executed for every sample that enters the processor. Obviously, low sample rates are desired for this reason. However, the sample rate requirement is dictated primarily by the information bandwidth of the signal. The information bandwidth in radio systems ranges from under 4 KHz for HF voice band channels to over 1 MHz for cellular systems. Spread spectrum (or CDMA) systems are a notable challenge, especially for military systems where interference excision techniques, or chip waveshaping (for LPI enhancement), are applied directly to the spread waveform. This requires that complex signal processing be applied at the chip level, which can be one or two orders of magnitude wider bandwidth than the information signal.

A well designed system will use a variety of sampling rates to achieve an efficient flow of data through the processor. At the A/D or the D/A stage, over sampling is quite often used. Over sampling of the signal is useful to shift aliases out of band and simplify filtering, so faster sample rates and narrower bandwidths are used. On the other hand, novel under sampling techniques are possible with stable, linear analog-to-digital converters. Under sampling techniques can be used to implement bandpass sampling - digitizing the signal in the second or third Nyquist zone, so that the desired signals will be aliased inband by the sampling. Both of these techniques can be combined as needed within the various stages of the software radio to enhance the signal to noise ratio yet maintain the sample rate at the lowest practical level.

When the time between samples is on the order of tens of microseconds to hundreds of nanoseconds such single-sample operations require hundreds of MIPS (million instructions per second) and/or MFLOPS (million floating operations per second) to Giga-FLOPS. A good FIR/IIR channel selection filter could require about 100 operations per sample at 30 Msps, or 3000 MIPS. Using a naive brute force approach, we would require 15 to 60 DSPs cooperating

for this section alone, repeated for every channel. As a result, even with faster devices, software on DSPs still cannot be used for the down conversion itself, but must still essentially operate at baseband (albeit a much wider baseband - up to a few MHz). Even the most fundamental demodulation or tuning algorithm requires 10 operations per sample, which would limit a DSP microprocessor to filtering signals with a bandwidth of a few hundred kHz. In a conventional voice-band cellular system, baseband processing requirements can range from 10 to 100 MIPS/MFLOPS per channel; while any digital signal processing at the IF frequency can drive the processing requirements to 500 MIPS/ MFLOPS and upwards of 10 GFLOPS [5].

We contend with these formidable processing challenges by abandoning the use of general purpose processors in favor of a mixed approach in which high speed digital hardware is used in the earliest stages, doing much of the filtering and processing in fast digital logic. When the signal reaches the post-IF stages, the processing load has been reduced considerable so that it can now be effectively handled by general purpose DSP processors. As long as this specialized hardware is versatile and is controllable to some extent from software, a hybrid architecture will meets our requirements. Most IF processing and chip rate processing can be off-loaded to these special purpose devices until the day that general purpose processors with sufficient processing power are available and cost effective.

2.3 DSP Hardware Alternatives

The most significant limiting factor in development of software radio systems has been the lack of sufficiently fast hardware - most notably, fast DSP microprocessors. As high performance, high speed ADCs have become available commercially, hybrid techniques using specialized digital hardware have become more common, while use of DSP microprocessors has lagged behind [5]. DSPs are getting ever faster, but it will be a while before we can use a single 'ultimate' chip to do everything. Instead the idea of using multiprocessing to share the effort seems attractive.

Multiprocessing as an alternative to the processing limitations of conventional DSPs can have only limited success. First of all, traditional DSP architectures were not well suited to multiprocessing. In fact, there are only one or two commercially available DSP processors which have the architecture to efficiently support multiprocessing - most notably the Texas Instruments TMS320C40. Also, software to support parallel and multiprocessing is scarce and expensive. Secondly, it is a characteristic of a DSP (as contrasted with a conventional microprocessor) that it must operate on a continuous flow of data. There are few functions in the software radio that could benefit from the power of parallel processors.

Software radios ideally place most IF, and all baseband, bit stream and source processing in a single processor. However, when we examine the speed requirements of the IF stage,

especially when spread spectrum is used, we conclude that we need a special purpose device - and this is where FPGAs come into favor. Some of the lower data rate anti-jam tactical communications standards, such as Have Quick and SINCGARS, are best implemented using high dynamic range software-oriented digital signal processing. In these radios, FPGAs could effectively provide the core of real-time sample rate and baud rate pipelined processing. They could also be used in their more conventional role of providing gate level support for the other processors and ASICs that make up the system [4].

Recent technical history has suggested that only software - and not hardware - possesses the programmability that is needed for versatile multi-role radio designs. The programmability of software is high; but the throughput necessary for any respectable data rate is low, making it suitable only for voice processing data rates. However, the availability of high speed FPGAs provides a greatly enhanced DSP capability which can be reprogrammed to handle wideband digital signal processing tasks. This permits a flexible architecture consisting of dedicated wideband ASICs, FPGAs, and programmable narrowband DSP processors. In the near future, reconfigurable modem architectures will provide in excess of 400,000 gates of programmable hardware with throughputs measured in the 100 millions of operations per second and at power consumption levels under 2 watts [4].

3. DSP MICROPROCESSORS

A modern programmable DSP microprocessor typically provides up to 200 MIPS or 50 MFLOPS. For example, the TMS320C40 has 50 MFLOPS at 25 MIPS with a 50 MHz clock.. There are many high performance DSP processors on the market, but they are not suited to all DSP applications. Their general purpose architecture makes these DSP processors flexible but they may not be fast enough or cost effective for all systems. The DSP processor provides flexibility through software instruction decoding and execution while providing high performance arithmetic components such as a fast array multiplier and multiple memory banks to increase data throughput. The performance limit for commercially available DSP processors currently tops out at about 50 MIPS [6].

Before exploring how DSP functions can be implemented on a variety of programmable logic devices, a broader definition of digital signal processing is in order. The term "DSP" applies broadly to discrete-time mathematical processes executed in real-time. These include such functions such as:

- Digital Filtering (FIR and IIR)
- Convolution

- Correlation
- Fast Fourier Transforms

Implementation of these functions involves only the basic digital operations of addition, multiplication and delay/shift as is indicated from the equation below:

$$y(n) = \sum_{k=0}^{N-1} h(k)\, x(n-k)$$

where x(n) can be interpreted as the input data sequence, and h(k) is the impulse response sequence of length N, and y(n) is the output. Depending on the data format and suitable choice of tap coefficients, a number of different functions result:

- Digital Filtering and Convolution - h(k) are the filter coefficients
- Correlation - h(k) refers to another input sequence
- Fourier Transform - h(k) are constants in complex exponential form

Most of these functions require the incoming data to be multiplied or added with various internal feedback mechanisms to perform the desired mathematical function. This primitive function which is so common to DSP algorithms is called the multiply/accumulate (MAC) [3]. The MAC may actually consist of 6 to 12 operations; however, to increase performance, most general-purpose DSP processors perform a MAC in a single clock cycle or less. Most DSP processors have a fixed-point MAC while some have a more expensive floating point MAC. Each tap of a digital filter requires one MAC cycle - for example a 16-tap filter requires 16 MAC cycles. Because most DSPs only have a signal MAC unit, each tap is processed sequentially, and all taps are processed during a single sample time interval, slowing overall system performance. Thus a 50 MHz (25 MIPS) DSP processor performs at less than 2 Msps [1].

The need to process instructions sequentially will always remain a fundamental performance limitation of microprocessors. Acceleration via dedicated hardware has long been a solution to this problem. Traditionally, this meant dedicated hardware in the form of an ASIC, or in some special cases, multiprocessing. Recently another viable alternative has been introduced - the Field Programmable Gate Array. The FPGA offers the advantage of fast hardware which can be reconfigured under software control. The use of FPGAs in DSP applications is the subject of the next section.

4. FIELD PROGRAMMABLE GATE ARRAYS

Programmable hardware has been available for many years - conventional memory devices are the most obvious example. Various PLDs (programmable logic devices) have long been used in implementing state machines and "glue" logic, among other things. However, the available devices have tended to have restricted architectures and to be rather small [7]. The last decade has seen a significant change with the introduction of a variety of field programmable gate arrays, as well as an evolution of some PLDs into much larger devices with extended architectures. Essentially, the FPGA is a general purpose programmable logic device consisting of a regular array of cells with distributed routing that can be configured with a specific design by the user, without the need to fabricate an application specific device (i.e., an ASIC) [8].

4.1 Programmable Logic Technology

There are a variety of FPGA architectures available, depending upon the manufacturer. However, there is one broad distinction that can be made regarding FPGA structure: the architectures are either *course-grained* or *fine-grained* [7]. Earlier devices were simple arrays of logic gates which were programmable in the field in much the same way as a conventional ROM. These devices are considered fine-gained in the sense that there can be a large number of very simple logic operations which can be interconnected. On the other hand, modern FPGAs have a relatively smaller number of more complex logic cells available.

Other than granularity, FPGAs are differentiated by their chip level architecture and their interchip wiring organization. As an example, the Xilinx 3000 family FPGAs consist of an array of cells called CLBs (configurable logic blocks). Each CLB contains two latches and a function generator as illustrated in Figure 2. The internal connections within the cell and the lookup table in the function generators are determined by configuration bits held in an integrated SRAM. This allows an individual cell to implement quite complex combinational and sequential functions. The routing resources allow the cells to be connected as required, at least in principle. In practice, the problem of routing a congested design is the major obstacle in obtaining highest performance.

FPGAs are just beginning to have a significant impact, although their cost is still relatively high (i.e., hundreds of dollars for the largest devices). Two application areas which traditionally have dominated their use are general purpose gate-level logic support (i.e., glue logic) and emulation of new IC designs. However, FPGA manufacturers believe that their products will change the way in which digital design is approached in a revolution similar to that engendered by the microprocessor [7]. The fact that FPGAs are now being investigated for use

in high speed DSP applications is an indication of the broad impact they may have in digital applications of all kinds.

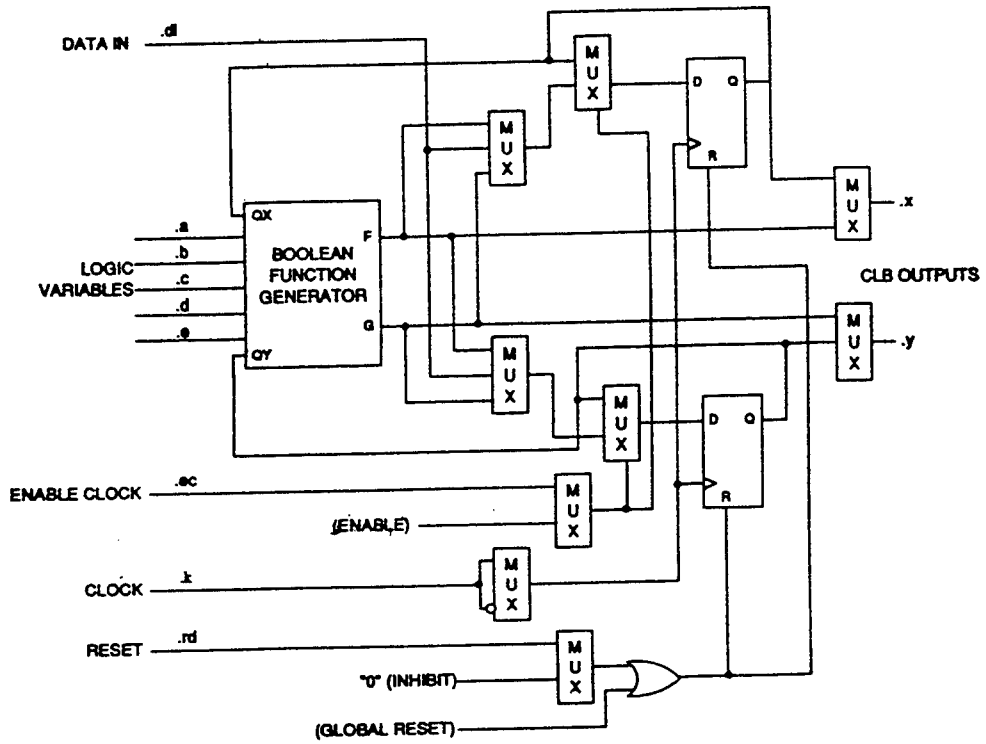


Figure 2 - Configurable Logic Block of the X3000 FPGA

4.2 Practical Consideration in the Use of FPGAs

Because the FPGA is programmable in manner similar to a microprocessor, it is already becoming widely used. However, the configuring of hardware to fit a specific computation is significantly different from the programming of a microprocessor. In particular, the microprocessor has a fixed instruction set, and all solutions are algorithmic in nature. In contrast, an FPGA's internal structure must be customized to implement a particular algorithm. Since digital hardware designs are not software driven, the overhead associated with command interpretation, scheduling and execution is eliminated and there is a substantial gain in speed. Furthermore, a hardware design can take advantage of parallel implementations to eliminate bottlenecks [2]. It is interesting to note that we may even combine the two approaches and compile a specialized microprocessor into the FPGA with a restricted instruction set chosen to suit any particular application.

It often occurs that a computation is better suited for either dedicated hardware or microprocessor software This is the situation we are examining in the software radio - when to use FPGAs and when to use DSP microprocessors. Simply stated, an FPGA is appropriate when

the design calls for the performance of an ASIC and the flexibility of a microprocessor. An FPGA should not be used if the algorithms to be implemented are complex, or vary significantly in structure or complexity. Determining when to offload DSP algorithms to FPGAs requires an analysis of speed versus problem size. At one end of the scale, problem size gets very large and direct hardware solutions become too difficult and expensive to build [2].

The advantage of FPGAs is that they represent a compact integrated programmable hardware solution which can be user-configured for any conceivable logic design. Current designs contain in excess of 40,000 logic gates, all under the control of the designer. On the other hand, FPGAs have some notable disadvantages. First their internal routing contributes substantial delay between logic elements resulting in a significant limitation in performance, although parallelism and pipelining can still be used. The second disadvantage is that it is not possible to execute a variety of arithmetic operations within the logic resources available. Added to this is that the programming of FPGAs is difficult, especially when implementing DSP functions [9].

4.3 Using FPGAs for DSP Applications

The FPGA has recently generated interest for use in DSP systems because of its potential to implement an infinite variety of custom hardware solutions while maintaining the flexibility of a conventional programmable device [6]. Although DSP microprocessors have complete algorithm flexibility, their performance is limited because algorithms are implemented by sequential MAC operations, as previously described. Additionally, DSP microprocessors have an overhead for reading in the operands and writing the result through a single data port. Therefore, a DSP microprocessor may require at least four cycles (i.e., read, multiply, add and write) to perform the simplest of algorithms, resulting in 10 MIPS performance from a 40 MIPS processor [1].

Because DSP algorithms are optimally mapped to the device architecture, FPGA performance can significantly exceed DSP processor performance. For example, a DSP microprocessor can implement an 8-tap FIR filter at 5 Msps. An FPGA can implement the same FIR filter at 100 Msps [1]. FPGAs will never completely replace general purpose DSP microprocessors, however. Current generation programmable logic addresses only the fixed point DSP portion of the market. General purpose DSPs still dominate in floating point performance. Also, general purpose DSP processors utilize familiar software methods, while using programmable logic requires a completely different approach on the part of the DSP designer Implementing DSP functions in FPGAs provide the following advantages over conventional DSP hardware:

a. *Parallelism* - Using FPGAs can lead to significantly higher performance than a typical DSP processor for some applications.

b. *Efficiency* - An FPGA can be optimized for specific algorithms, thus achieving the performance of hardware with the flexibility of software.

c. *In-circuit reconfigurability* - Permits the algorithm or function to be changed while operating in-circuit. An additional benefit of FPGAs over ASICs is that they can be reprogrammed, on the fly, in the system. Consequently, a single FPGA can implement different DSP functions at various times in a system to boost overall performance.

d.. *Adaptability* - A device that can implement large internal RAM blocks can be used to implement real-time adaptive functions at a throughput that cannot be matched by conventional DSP solutions.

4.3 .1 Alternative Arithmetic Options for FPGA

The primary limitation of the FPGA when used in DSP applications is arithmetic - most notably, multiplication. A hardware multiplier is a reasonably complex circuit, as evidenced by the fact that conventional DSP microprocessors contain only a single hardware multiplier, and it occupies most of the real estate on the chip. A state-of-the-art FPGA can support no more than a handful of multipliers, meaning that brute force multiplication is to be avoided in some of the most common operations - e.g., filtering or correlation. The problem is, how does one built an FIR filter when multipliers cannot be used.

In order to understand the performance of the FPGA relative to the DSP processor, a comparison of FPGA multiplication alternatives and their performance relative to custom multiplier solutions is needed. A core operation in DSP algorithms is multiplication. Often the computational performance of a DSP system is limited by its multiplication performance, hence the multiplication rate of the system must be maximized. Custom hardware systems based on ASICs and DSP processors maximize multiplication performance by using fast parallel array multipliers either singly or in parallel.

When implementing multipliers in hardware, two basic alternatives are available: The fully parallel array multiplier and the fully bit-serial multiplier. The advantage of the fully parallel array multiplier is that all of the product bits are produced at once which generally results in a faster multiplication rate. The multiplication rate for this adder is simply the delay through the combinational logic. However, parallel multipliers also require a large amount of area to implement. Bit serial multipliers on the other hand generally require only 1/Nth the area of an equivalent parallel multiplier but take 2N bit times to compute the entire product (N is number of bits of multiplier precision) [6]

4.3.2 FPGA Applications in Software Radio Systems

The DSP Functions that FPGAs do best are those requiring high sample rates and short word length. They are especially suited for FIR filter designs employing lots of filter taps and fast correlators. The lookup table architecture of FPGAs provides a fast and efficient way to build correlators [3] More taps can be added to the parallel filter with only a small performance tradeoff with additional parallel silicon resources. In contrast, DSP processors exhibit a linear decrease in performance as the number of taps increases. (see Table 1). An 8-tap, 8-bit FIR filter implemented on an Altera device needs only 80% more silicon than one 8 x 9 bit fixed multiplier (Table 2)[1].

Table 1. Fully Parallel 8-Bit FIR Filter in FLEX 8000A (-2 Speed Grade Device)

Number of Taps	FLEX 8000A Performance -2 Speed Grade (MSPS)	Equivalent MIPS (DSP Processor)
8	104	832
16	101	1,616
24	103	2,472
32	105	3,360

Table 2. Silicon Resource Comparison

Function	Inputs & Outputs	FLEX 8000A Logic Cells
FIR filter	8-bit data, coefficients, 17-bit output	296
Fixed-point multiplier	8-bit × 9-bit data, 17-bit output	164

FPGAs can efficiently implement IIR filters. For example, a lookup table based vector multiplier can be used to create a complete second order section of an all pole analog filter. The vector multiplier requires the same resources and operates at the same speed as a fixed point multiplier. A Butterworth filter can run at a rate of 25 Msps and require only 139 logic cells. [1]

Altera has developed high speed FIR filter megafunctions that are optimized for their own FPGA structure. These filters can be implemented in parallel or serial form allowing a tradeoff between silicon resources and performance. Parallel filters can perform at rates up to 100 Msps enabling digital processing of RF-IF data. Serial filters require less logic and still perform at 5 to 6 Msps. In a Spread Spectrum RF modem application, an Altera FPGA can implement the receivers correlation filter function at a chip rate over 60 MHz. A DSP processor

can perform the remaining tasks, such as quadrature phase shift key (QPSK) demodulation. The resulting DSP application can deliver six times the data rate as the DSP processor alone.

4.4 Rapid Prototyping Concepts

Designing with FPGAs requires computer assistance at almost every stage of the design including detailed specification, simulation, placement, and routing. The use of schematic capture based CAD tools is a common approach to the design of custom logic devices using FPGAs. This process is often combined with logic level simulation to verify a specific design. One method of increasing the range of architectural solutions that a designer may explore in a reasonable time is to specify the DSP system with a hardware description language (HDL) [10]. This steps the design process up one level and allows a generic functional description of the target system which can be further simulated or implemented directly onto an FPGA after the HDL code is converted using the FPGA manufacturer's software.

In DSP applications, arithmetic circuitry for operations such as addition, subtraction and multiplication are commonly required. These arithmetic circuits can be designed and implemented by employing user-generated or manufacturer-provided sub-circuits, which can be reused. However, as these designs can only be simulated at the logic gate level, it is difficult to verify the functional performance of the algorithms being implemented. It is particularly difficult to determine the potential undesirable side effects of finite precision arithmetic, as this may require that large data sets be simulated and translated from numerical values to logic levels and vice versa [10]. However, new software tools are being developed which raise the design process to yet another level, allowing the designer to begin at the system level.

Simulation tools such as Cadence's Signal Processing Worksystem (SPW) now have features which allow the engineer to design hardware logic systems and DSP fixed point systems using the traditional block diagram functional description of the circuit. This design is then immediately converted into a hardware description language. Other SPW tools allow the design to be simulated via the HDL description of the system and then linked into a manufacturer's software tools which support specific devices. Most manufacturers, in the interest of making their product more attractive to their customers, have developed a set of stock logic elements which can be reused within their device to assist the engineer in quickly achieving any design.

Once suitable design tools and automatic methods are perfected, designers and programmers will be able to create custom hardware circuitry and pipelines to suit the problem at hand - the term 'soft hardware' suggests that hardware will be come as readily created and malleable as software. In a practical sense this will mean that the turn-around time for custom hardware will be just as short as software development is today

5. A DSP Testbed for Tactical Radio System Evaluation

Current research at the USAF Rome Laboratories is focused on the development, testing and evaluation of algorithms for future Air Force radio systems. It is understood that these are radio systems which are implemented using state of the art digital signal processing hardware. The question of how to make the best use of currently available DSP hardware is one that we have to answer. We have concluded that the FPGA is best used in two situations - in the post IF stage of the receiver before the spreading sequence is removed from the signal; and in the data bit stream (post-modem) stage of the receiver where such functions as de-interleaving, error correction decoding, source decoding and data decryption functions are used.

A proposed system architecture for the DSP testbed is shown in Figure 4. FPGAs are used to advantage in the chip stream section of the transmitter and receiver where a few simple DSP operations need to be performed at a high rate of speed. In this section interference excision filters of any description can be implemented, including FIR filters, Fast Fourier Transforms, and even adaptive filters. A need still exists for the general purpose DSP microprocessor to perform the complex operations of (possibly synchronous) demodulation, system timing, carrier extraction and adaptive equalization. Once the bit level decision has been made, digital signal processing is no longer needed - the signal of interest is now in the form of a binary information bit stream. This data bit stream can now be processed entirely in digital hardware. In this manner, the testbed makes appropriate use of both general purpose DSPs and FPGAs to provide the most effective and flexible testbed available with current DSP hardware technology.

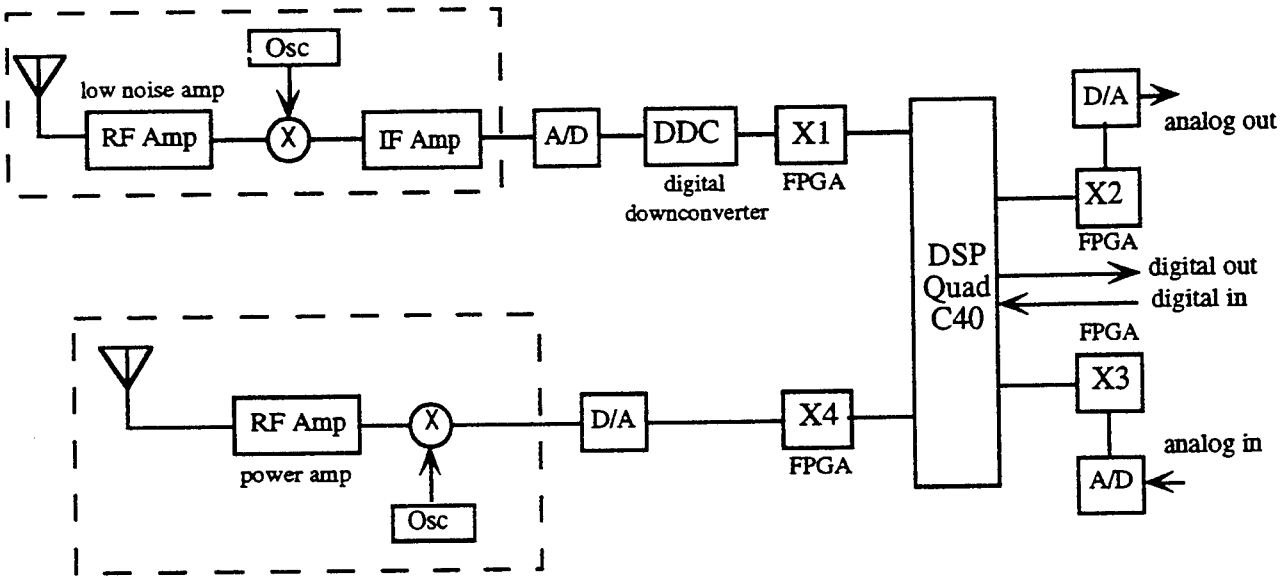


Figure 3 - Software Radio Testbed

6. Conclusions and Recommendations

This report has briefly examined the emerging field of FPGAs and their application to traditional DSP problems. We compared the computational power of the FPGA to that of the DSP processor, and we found that FPGAs are more suited to simple algorithms executed at high speed, while DSP microprocessors are better suited for slower, more complex tasks. DSP microprocessors are limited by their serial-based architecture, while FPGAs can implement any architecture which may be optimum for a particular algorithm. On the other hand, we recognized the shortcomings of the FPGA - these devices are essentially still in their infancy, from a technological standpoint, and there are problems to be solved by the manufacturers before FPGAs seriously challenge DSP microprocessors in the DSP market.

Designing with FPGAs is still a difficult task. The software necessary for efficient FPGA design is extremely expensive and rare. Often, a considerable amount of gate-level modification is necessary on the part of the engineer to realize the performance that the FPGA is capable of delivering. The current cost of FPGAs is an indication of their novelty in the marketplace. Prices will need to drop by several orders of magnitude before FPGAs will be a common commodity in the DSP designers bag of tricks.

Given the state of current FPGA technology, we examined how best to use these devices in the emerging area of software radio, and suggested a testbed that would be suitable for investigating software radio algorithms. We found that applications at the chip level were suitable for the speed of the FPGA as long as processing at this stage can be kept relatively simple. Filters, correlators and fast Fourier transformers are examples of processing suitable for this stage. After the spreading sequence has been removed, DSP microprocessors are the preferred solution for demodulation, timing and carrier recovery operations. Once the bit level decision has been made, processing of the bit stream with error correction and source compression algorithms makes the FPGA a suitable alternative once again. Therefore, in this testbed the FPGA is used as both a DSP processor and as a state machine. For the testbed described in this report, use of the FPGA is ideal

It is undeniable that the real impact of FPGAs is yet to be felt. However, it is also undeniable that restructurable hardware such as the FPGA will have a major impact on digital signal processing in the near future.

7. References

[1] Digital Signal Processing in FLEX Devices, ALTERA Product Information Bulletin 23. See *http://www.altera.com*.

[2] Leo Petropoulos, Replace Digital Signal Processors with HCPLDs, *Electronic Design*, September 5, 1995, pp. 99 - 104.

[3] Steven Knapp, *Using Programmable Logic to Accelerate DSP Functions*, Xilinx Application Note, 1995.

[4] Eric. L. Upton and Thomas J. Kolze, "Reconfigurable Modems Serve as Multi-Application Communications Node Integrators" 1993 Conference of the American Institute of Aeronautics and Astronautics, pp 1 - 13.

[5] Rupert Baines, "The DSP Bottleneck," *IEEE Communications Magazine*, May 1995 pp. 46 - 54.

[6] Russell Petersen and Brad Hutchings, "An Assessment of the Suitability of FPGA Based Systems for Use in Digital Signal Processing," *Field Programmable Logic and Applications Proceedings of the 5th International Workshop,* FPL-95, Oxford, United Kingdom, August/September 1995, pp. 293 - 302.

[7] A. Lawrence, A. Kay, W. Luk, T. Nomura, "Using ReconfigurableHardware to Speed up Product Development and Performance," *Field Programmable Logic and Applications Proceedings of the 5th International Workshop,* FPL-95, Oxford, United Kingdom, August/September 1995, pp.111 - 117.

[8] S. Kotta and S. Simanapalli, "Rapid Prototyping of a Digital Signal Processor," "Rapid Harware Prototyping of Digital Signal Processing Systems using Field Programmable Gate Arrays," *Field Programmable Logic and Applications Proceedings of the 5th International Workshop,* FPL-95, Oxford, United Kingdom, August/September 1995, pp.844 - 847.

[9] Paul Dunn, "A Configurable Logic Processor for Machine Vision,"*Field Programmable Logic and Applications Proceedings of the 5th International Workshop,* FPL-95, Oxford, United Kingdom, August/September 1995, pp. 68 - 77.

[10] L.E. Turner and P.J.W Graumann, "Rapid Harware Prototyping of Digital Signal Processing Systems using Field Programmable Gate Arrays," *Field Programmable Logic and Applications Proceedings of the 5th International Workshop,* FPL-95, Oxford, United Kingdom, August/September 1995, pp.129-138.

[11] Joe Mitola, The Software Radio Architecture, *IEEE Communications Magazine*, May 1995, pp. 26 - 38.

A NETWORK FLOW HEURISTIC FOR GRAPH MAPPING

Mark Purtill
Assistant Professor
Department of Mathematics

Texas A&M University-Kingsville
Campus Box 172
Kingsville TX 78363

Final Report for
Summer Faculty Research Program
Rome Laboratory

Sponsored by
Air Force Office of Scientific Research
Bolling AFB, Washington DC

and

Rome Laboratory, Rome NY

September 1996

A NETWORK FLOW HEURISTIC FOR GRAPH MAPPING

Mark Purtill
Assistant Professor
Department of Mathematics
Texas A&M University–Kingsville

Abstract

We introduce a network-flow based approach to partitioning computation which can be modeled as a graph for parallel computers. This includes several problems in computational physics. We discuss some applications and ongoing work on implementing the algorithm.

A NETWORK FLOW HEURISTIC FOR GRAPH MAPPING

Mark Purtill

1 Introduction

Many problems in computational physics can be viewed as computations carried out on graphs (networks); one such problem arises in particle-in-cell (PIC) plasma simulations. As many multiprocessors (massively parallel computers) can also be viewed as graphs, the problem of partitioning such problems to run efficiently on a multiprocessor is a graph theoretic problem. It can be viewed as choosing a graph homomorphism between the two graphs minimizing a function representing balance and communications requirements. As computational requirements change, we must "rebalance"; this amounts to choosing a new homomorphism "close to" the previous map.

Unfortunately, this problem is NP-complete, which means it is very unlikely that we can find an efficient, exact algorithm. Therefore, we must rely on heuristic methods. Many such methods have been proposed for this and related problems.

We propose a network flow based algorithm for the rebalancing problem. Basically, we: find the imbalances (a global step in some cases); use a network flow (min-cost flow) algorithm to find transfer to balance the work; and use a modified Kernighan-Lin algorithm [16] to transfer vertices.

To test this and other approaches, we have developed a test bed program which runs on both networks of workstations and on multiprocessors. The test bed is written in C++, using modern programming methods such as object-oriented design and literate programming.

This report is organized as follows: the next section is a summary of some graph theoretic concepts. In section 3 we discuss the problem. The section after that describes the proposed algorithm, and compares it to some other algorithms which have been proposed. In the last sections we discuss potential applications and experimental work.

2 Graph Theory[1]

A *graph* G consists of a set of objects, called *vertices* (singular *vertex*), denoted $V(G)$ and a set of unordered pairs of vertices, called *edges*, denoted $E(G)$. (For our purposes, both of these sets must be finite.)

The two vertices v, w in an edge are called its *ends*, and the edge is said to *connect* its ends, which are said to be *adjacent*; we write $v \underset{G}{\sim} w$ ($v \sim w$ if G is clear from the context). We do not allow an edge to connect a vertex to itself, nor for more than one edge to connect a given pair of vertices (that is, our graphs are "simple").

[1] This section is taken from [22].

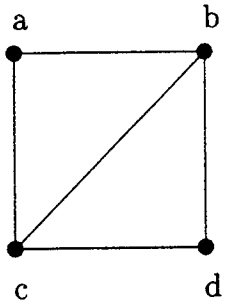


Figure 1: A drawing of a graph; the disks represent vertices, and the lines represent edges.

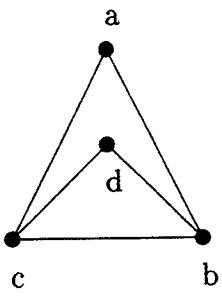


Figure 2: A different drawing of the same graph as in figure 1.

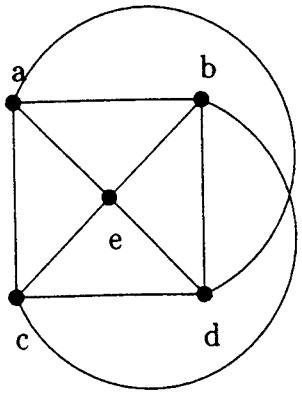


Figure 3: A drawing of a graph which cannot be drawn without edges crossing: the complete graph on five vertices.

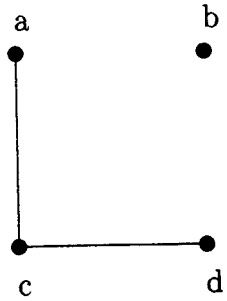


Figure 4: This graph is disconnected because (*e.g.*) there is no path from vertex a to vertex b.

A graph may be represented by a drawing (picture), such as figure 1; however, the geometric information (such as coordinates of the dot representing a vertex) are *not* part of the graph; figure 2 represents the same graph G as figure 1: the vertex set is $V(G) = \{a, b, c, d\}$ and the edge set $E(G) = \{\{a, b\}, \{a, c\}, \{b, c\}, \{b, d\}, \{c, d\}\}$.

If every pair of vertices is an edge (that is, all pairs of vertices are connected), then the graph is called "complete". Figure 3 is a complete graph on five vertices.

Note that in some cases (such as figure 3), it is necessary for the edges of the drawing to cross; this has no effect on the graph itself. Graphs that can be drawn without crossings are called planar graphs; these have various nice properties (see *e.g.* [8]). We do not assume our graphs are planar, since many interesting graphs are not planar.

A sequence of vertices $(v_1, v_2, v_3, \ldots, v_n)$ such that each pair of vertices v_i, v_{i+1} is called a *path*. For instance, in the graph in figure 1, (a, c, b) is a path, while (a, d, b) is not. (Note that we can also specify a path by listing the edges in that path rather than the vertices.) The *length* of a path is the number of *edges*

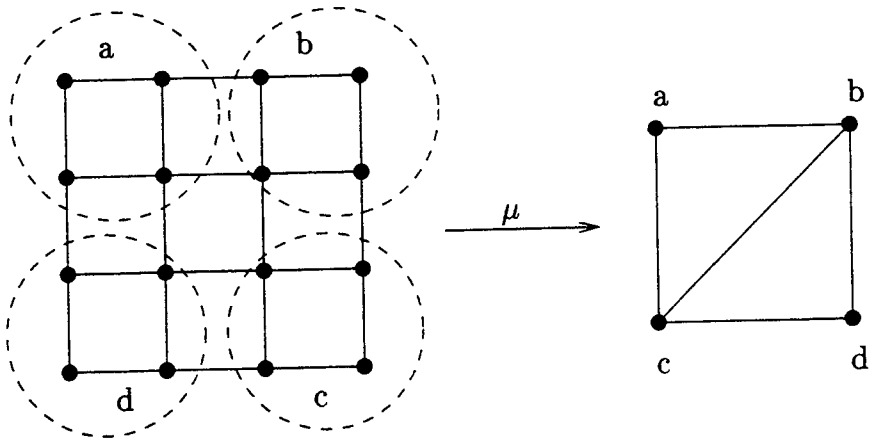


Figure 5: A graph map; each of the dashed circles in the graph on the left hold the vertices in the preimage of the vertex in the graph on the right with the same label.

in it.

If there is a path between any pair of vertices in the graph, then the graph is called *connected.* The graphs in figures 1, 2 and 3 are connected; that in figure 4 is not connected. All graphs considered in this paper will be connected. The *distance* between two vertices in a graph is the length of the shortest path connecting them; the *diameter* of the graph is the longest distance between two vertices. For instance, the graph in figure 1 has diameter 2.

A *graph map* μ from a graph G to a graph H, denoted $\mu\colon G \to H$ is a function μ from the vertex set of G, $V(G)$, to the vertex set of H, $V(H)$, so that for every edge $\{v,w\}$ of G, either $\mu(v) = \mu(w)$ or $\{\mu(v),\mu(w)\}$ is an edge of H; that is:

$$\text{if } \{v,w\} \in V(G), \text{ then either } \{\mu(v),\mu(w)\} \in V(H) \text{ or } \mu(v) = \mu(w)\,.$$

(Aside for graph theorists: if we assume each vertex of each graph is equipped with a self-loop, this coincides with the usual definition of *e.g.* [8].)

For each vertex v of H, we denote by $\mu^{-1}(v)$ the set of all vertices of G which map under μ to the vertex v (the *preimage* of v); that is $\mu^{-1}(v) := \{\, w \in V(G) \mid \mu(w) = v\}$. A sample graph map is shown in figure 5.

The *valance* or *degree* of a vertex is the number of edges of which it is an endpoint, or, equivalently, the number of vertices which are connected to it by an edge. In figure 1, vertices a and d have valance 2, which b and d have valance 3. If all of the vertices of a graph have the same valance, the graph is called *regular*; for instance, figure 3 shows a regular graph (of valance 4).

For a simple application of the graph concept, consider a multiprocessor; that is, computer consisting of a number of different processors, each with its own memory, which can communicate with each other *via* communications links. Assume each link connects exactly two processors.

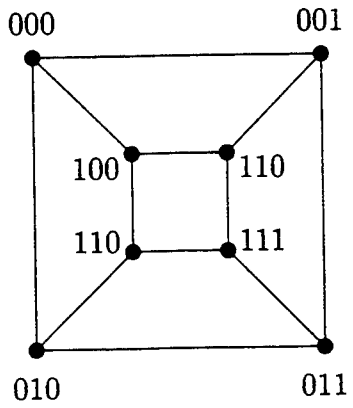


Figure 6: A three-dimensional hypercube graph.

This arrangement of processors can be expressed as a graph; the vertex set of the graph is the set of processors, while the edge set is all the pairs of processors which are directly connected. (So the set of edges is the same as the set of links.) When we think of a graph as representing a multiprocessor, we will refer to the vertices (edges) as processors (links).

A path $(p_1, p_2, \ldots, p_n)$ in this graph is a way that a message could be passed from processor p_1 to processor p_n through processors p_2 through p_{n-1} *via* direct links, and the graph is connected precisely when each processor can communicate with each other processor (directly or indirectly).

For example, a common network connection is the "n-dimensional hypercube", which consists of 2^n processors, labeled 0 to 2^{n-1}. Two processors are connected if the binary representation of their labels differ in exactly one bit. See, for example, figure 6. For instance, the computer I did some work on this problem in 1994 (an Intel iPSC/860 called `neutrino`—see [21]) had the topology of the four-dimensional hypercube.

Often it is handy to be able to associate a number with each vertex or each edge of the graph; we call such numbers *weights*. Vertex weights are numbers associated with each vertex of the graph, while edge weights are numbers associated with each edge. For instance, we might have vertex weights giving the relative speed of each processor, and edge weights giving the speed or bandwidth of each communications link.

Formally, if we say a graph G has "vertex weights w", we mean w is a function from the vertex set $V(G)$ of G to the real numbers, and write $w(v)$ for the weight of vertex v. Similarly, if we say G has "edge weights u", we mean u is a function from the edge set $E(G)$ of G to the real numbers. In this paper, we will only need vertex and edge weights which have non-negative values.

A note on notation: We use a variant of Iverson's convention (see Knuth [17]), and define $[S]$ to be 1 if the statement S is true and 0 otherwise. In sums, this 0 is especially strong; thus, for instance, $\sum_i [1 \le i \le n] \frac{1}{i} = \sum_{i=i} n \frac{1}{i}$.

3 The Problem

Let G and H be two connected graphs. The graph G is called the "guest graph", and the graph H is called the "host graph", and we assume that $\#V(G) >> \#V(H)$. Let $w\colon V(G) \to \mathbf{R}^{\oplus}$, $c\colon E(G) \to \mathbf{R}^{\oplus}$, $p\colon V(H) \to \mathbf{R}^{\oplus}$ and $b\colon E(H) \to \mathbf{R}^{\oplus}$ be functions giving non-negative weights to the vertices and edges of G and H.

Suppose the graph G represents the communications pattern of a computation: Each vertex v represents some work to be done, with $w(v)$ representing the amount of work at vertex v. Each edge $e = (u, v)$ indicates that processes u and v must communicate during the work: $c(e) = c(u, v) = c(v, u)$ represents the amount of communications.

More specifically, we assume the graph G is working in an iterative way: at each time step t, a vertex computes some quantities (say $q(v, t)$) using the outputs of its neighbors at the previous step. Thus $w(v)$ represents the amount of effort needed to compute $q(v, t)$, and $c(u, v)$ the amount of effort needed to send $q(v, t)$ to u and $q(u, t)$ to v.

Now, the graph H is thought of as a multi-processor or parallel computer: Each vertex v represents a processor with processing power $p(v)$. Each edge $\epsilon = (\upsilon, \nu)$ represents a bidirectional communications link with bandwidth $b(\epsilon)$ connecting the processors υ and ν. (From now on "vertex" and "edge" will refer to G; we will speak of "processors" and "links" in H; similarly, the vertices and edges of G will be denoted with Roman letters, while those of H will be denoted with Greek letters.)

For special graphs G or H, this problem reduces to various other problems which have been studied before. For instance, if G is a collection of disconnected vertices (no edges), then we have a scheduling problem on graphs [3, 5, 14]. On the other hand, if $H = K_2$ (the graph with two vertices connected by an edge), then we have the well-studies minimal graph bisection problem [1, 11, 16].

3.1 The Mapping Problem

The "mapping problem" is to find an assignment of vertices of G to processors of H so as to minimize the running time of the algorithm. More specifically, we wish to find a graph G' (obtained from G by subdividing some edges) and a map $i\colon G' \to H$ minimizing some combination of the following quantities:

- Some measure of the difference between G and G', such as the length of the longest path in G' corresponding to a since edge of G (which is called the "dilation");
- The work done at the most loaded processor of H: $W = \max_{v \in V(H)} \dfrac{w(i^{-1}(v))}{p(v)}$.

  (Here $i^{-1}(v) = \{ v \in V(G') \mid i(v) = v \}$, and we extend w to G' by assigning 0 work to vertexes added in subdivision, and for a set of vertices $V \subset V(G)$, let $w(V) = \sum_{v \in V} w(v)$.)

- The weight of edges of G' crossing the most loaded link of H:

$$C = \max_{v \underset{H}{\sim} \nu} \sum_{u \underset{G}{\sim} v} [i(u) = v][i(v) = \nu] \frac{c(u,v)}{b(v,\nu)}.$$

  (Here c has been extended from G to G' by setting $c(e) = c(f)$ if f is an edge of G' obtained by subdividing and edge e of G.

The form of the combination (and the exact form of these terms) depend on the exact model of parallel computation assumed for H. (For instance, in the above we are implicitly assuming that no computation (at a processor) is required to pass a message.)

3.2 The Remapping Problem

In the mapping problem, it was assumed that the work $w(v)$ at each vertex v of G was constant for the life of the computation (and similarly for c, p, and b). The "remapping problem" is to take an additional weight function $s\colon V(G) \to \mathbf{R}^{\oplus}$, and an "old" mapping $i_0\colon G' \to H$ (possibly a solution to the mapping problem for some w_0, c_0, p_0, b_0), and find a new graph G'' and new graph homomorphism $i\colon G'' \to H$, where G'' is obtained from G as before (which means some of the subdivided edges of G in G' may be recollapsed, and some other edges may be subdivided), and in addition to the quantities minimized in the mapping problem, we also consider:

- The size of the vertices of G which must be moved between processes (again across the most loaded link): $X = \max_{v \underset{H}{\sim} \nu} \sum_{v} [v \in i_0^{-1}(v)][v \in i^{-1}(\nu)] \frac{s(v)}{b(v,\nu)}$.

By selecting a distinguished processor $v \in V(H)$, setting $i_0(v) = v$ for all $v \in V(G)$, and neglecting X in the minimization, the mapping problem can be reduced to the remapping problem.

3.3 Complexity of Remapping

Most problems of the form of the mapping and remapping problems (that is, with the details filled in different ways) are NP-hard. For example, considering $H = K_2$, setting $p(v) = 1$ for both vertices of H, and considering only minimizing the maximum weight of work on any processor W gives the set partition problem [7].

Reductions of several other problems in Garey and Johnson [7] are easy to construct.

Because of this, work on solving this problem has focused on heuristics.

4 Solving the Problem

4.1 The Flow Algorithm

One step in the flow algorithm is as follows:

- Find the imbalances (a global step in some cases);
- Use a network flow (min-cost flow) algorithm to find transfer to balance the work;
- Use a modified Kernighan-Lin algorithm [16] to transfer the vertices.

We will discuss each of these steps in turn. In some cases, there may be an additional step; for instance, after finding imbalance in step 1, we may wish to consider whether given the expected time remaining in the computation, it is worth the effort to rebalance. (If the computation is near completion, the time saved by balancing may not make up for the time it costs.)

4.1.1 Finding Imbalances

In general, finding imbalances will require global communications. Many modern parallel computers have efficient global communications routines which will make this step more efficient. Also, we may be able to carry out this step in parallel with computation (at the cost of not having the most up-to-date knowledge of the load). Even so, this step may not scale well.

Fortunately, in some cases the global communications may be much reduced or eliminated entirely. For instance, if the total amount of work in the system, and the relative speeds of the processors is know (for instance, if all the processors are equally fast), the local computation suffices to determine whether a processor is overloaded or underloaded. (Each processor can compute its "fair share" of the know amount of work in the system, and compares it to the work it is doing).

4.1.2 Using min-cost flow

We set up the flow problem as follows: the graph the flow will take place on is the host graph H, that is, the graph of the parallel processor. To this, we adjoin a source node σ, which is connected to every processor which is overloaded; and a sink node τ, which is connected to every processor which is underloaded.

Each edge of H, say (μ, ν) is given infinite capacity ($cap(\mu, \nu) = \infty$) and a cost of 1 unit per unit flow ($cost(\mu, \nu) = 1$). The edges connected σ and τ to processors are given capacity equal to the overload or underload at that vertex, and a cost of 0. (A cost of 1 would work as well, since the max-flow will saturate all of these edge.) Thus, if (σ, μ) is an edge of $\tilde{H} = H \cup \{\sigma, \tau\}$, then $cap(\sigma, \mu)$ = the overload of processor μ and similarly for underloaded processors. We assume $cap(\mu, \nu) = cap(\nu, \mu)$ and $cost(\mu, \nu) = cost(\nu, \mu)$.

A *flow* on the network $\tilde{H}$ is an assignment of a real value $f(\mu, \nu)$ to each pair of processors (μ, ν) connected by an edge in $\tilde{H}$ which satisfies the following:

- for all vertices μ and ν of $\tilde{H}$ $f(\mu, \nu) = -f(\nu, \mu)$;
- for all vertices μ of $\tilde{H}$, $f(\sigma, \mu) \geq 0 \geq f(\tau, \mu)$ (that is, flow only leaves the source σ and only enters the sink τ);
- for all vertices μ and ν, $|f(\mu, \nu)| \leq cap(\mu, \nu)$ (that is, the flow does not exceed the capacity of the edges); and

- for all processors μ of H (that is, not including σ and τ), $\sum_\nu [\mu \sim \nu] f(\mu, \nu) = 0$ (that is, the amount flowing into a processor must be the same as the amount flowing out). (Note that in the sum, ν may equal σ and τ.)

The *value* of a flow f is $\sum [\mu \sim \sigma] f(\sigma, \mu) = \sum [\tau \sim \mu] f(\mu, \tau)$; this is the total flow thru the network from σ to τ. A flow f is said to be a maximum-flow or just a *max-flow* if there are no flows with higher value.

Thus, a flow of $\tilde{H}$, by neglecting the flows from σ and to τ, gives a transfer of work from processors in H to *adjacent* processors in H. Moreover, a max-flow gives such a transfer which, if it could be carried out, would result in all processors having equal work.

Proof. First, by the max-flow/min-cut theorem (see, for instance, [6]), it is easy to see that all edges going from σ or to τ will be saturated, that is, $f(\sigma, \mu) = cap(\sigma, \mu)$ for all overloaded processors μ and $f(\mu, \tau) = cap(\mu, \tau)$ for all underloaded processors μ. (This is because the only two min-cuts are the one consisting of all of the edges from σ, and the one consisting of all the edges to τ.)

Now, consider an overloaded vertex $\mu \in V(H)$. The total flow in $\tilde{H}$ from μ is 0. The flow from σ to μ is the overload of μ, $cap(\sigma, \mu)$, by the previous paragraph, so the remaining flow must sum to $-cap(\sigma, \mu)$, and if such a transfer were made, the overload of μ would be exactly eliminated. Similarly, for nodes with an underload, the underload is exactly eliminated. □

The cost of a flow f is $\sum [(\mu, \nu) \in E(\tilde{H})] f(\mu, \nu) cost(\mu, \nu)$; a *min-cost max-flow* is a max-flow for which there is no cheaper max-flow. (There may be cheaper flows; just not cheaper *max*-flows.) A min-cost max-flow on $\tilde{H}$, then, gives a transfer which eliminates the imbalance in H with the fewest total transfers.

There are many algorithms for finding a min-cost max-flow in a network, including both serial and parallel algorithms. If the previous step (finding the imbalances) requires non-scaling global communications, it would probably be best to have one processor compute the flow with a serial algorithm and send the flow to the affected processors. In this case, a processor would send the amount of work it is doing to a designated processor, then receive a reply saying "send w_1 units of work to processor p_1, *etc.*".)

On the other hand, if finding the imbalance can be done entirely with local computation, using one of the parallel algorithms for finding the flow would be more scalable. (For small problems, of course, having one processor compute the flow, as in the previous case, would also work.)

4.1.3 Transferring Vertices of G

Once the local flow values are obtained, we must transfer vertices of the computation graph G between neighboring processors in H so as to realize the flow as well as possible. It may not be possible to realize the flow exactly, depending on the flow requested and the weights of the vertices available to be moved. (Even if it is possible, we would not necessarily want to realize the flow exactly; the overall cost of communications between the processors must be taken into account as well.)

To do the transfer of work, we use a modification of the Kernighan-Lin algorithm for locally improving a graph bisection [16]. The main alteration is that rather than the processors alternating in sending nodes across, whichever processor needs to transfer vertices away in order to more closely realize the flow sends

the next node. Also, the function for determining the "best" exchange could be altered to take into account imbalance as well as cost of communications between the processes.

Suppose, then, we have nodes V_0 of G on processor μ_0 of H, nodes V_1 on processor μ_1, and we need to transfer w units of work from μ_0 to μ_1. Let $X_i = \varnothing$; $X_i \subset V_i$ will be the vertices transfered from processor μ_i to processor μ_{1-i} to form the "current" bisection. Let w_i be the weight of the vertices in V_i.

Mare the current partition (V_0, V_1) as the "best-so-far", and loop over the following.

> Let μ_i be whichever processor needs to send work to the other (μ_0 if $w - w_0 + w_1 > 0$, and μ_1 otherwise), and let $\mu_j = \mu_{1-j}$ be the other processor.
>
> On μ_i, choose the vertex $v \in V_i - X_i$ so that the partition $(V_i - X_i \cup X_j - v, V_j - X_j \cup X_i \cup v)$ is "as good as possible". (This is the partition that would be obtained if v were added to X_i.) "As good as possible" means the least number (or weight) of edges between the two parts of the partition, along with some consideration of the total work transfered between the two processors compared with w. Using data structures such as those found in [4], v can be found in linear time.
>
> Now, whether or not it improves the partition, v is added to X_i. We compare the new partition to the partition marked as best so far. If the new partition is better than the old best so far, it takes over as best so far.
>
> This loop terminates when there are no more vertices to transfer in the required direction. At this point, nodes are transfered between the processors as given by the "best so far" (X_0, X_1): processor μ_i send the vertices of G in X_i to processor μ_{1-i}.

This procedure can be executed a fixed number of times, or until no further improvement is obtained.

4.2 Other Algorithms

Several methods for a similar problem are mentioned in [27], along with some experimental results. Below, we will discuss several proposed methods and compare them to our flow algorithm.

4.2.1 Local Averaging

One method which has been proposed for load balancing in graph-like parallel processors is for each processor to balance with its neighbors. One way of doing this would be to have each processor send the amount of work it has to each neighbor, then use this information to decide how much work it should send to each neighboring processor.

This method could be modified to solve our remapping problem; something like our modified Kernighan-Lin algorithm would be used to transfer vertices of G between processors of H.

Compared with our flow method, "local averaging" has the advantage that it never needs to do global communications. It is also somewhat simpler, in that it doesn't need the network flow algorithm.

On the other hand, local averaging may take a long time (or forever) to converge to equal loads, even if the loads are not changing [10]. Depending on granularity, the flow algorithm will get very close to balance in one (longer) step.

4.2.2 Diffusion

Diffusion algorithms, such as those found in [10, 19, 26], are similar to local averaging, but are cleverer about moving work to avoid, for instance, having an underloaded processor suddenly become overloaded because all its neighbors send it load.

These algorithms have the advantage of never needing global communications, while the flow algorithm may in some problems. In addition, like the flow algorithm, results about convergence can be proven.

However, these algorithms are quite complicated, even compared with network flow, and they appear to require that the host graph be a grid in some Euclidean vector space, such as the plane or three-space.

4.2.3 Static Partition Algorithms

In our setup, a static partition problem is a solution to the mapping problem with the host graph a complete graph. As noted above (in section 3.3), this problem is still NP-hard.

Numerous heuristics have been proposed, most for bisection (which in our notation is the case $H = K_2$, although repeated bisections can be used to handle $H = K_{2^n}$ or a hypercube). Some of these (in no particular order) are:

- the Kernighan-Lin algorithm [16], starting with a random partition; there are also some variations of interest [4, 23].
- the spectral method [11, 12, 20, 25];
- several multilevel methods [1, 13, 15];
- various other methods [2, 9, 18, 24].

5 Applications

5.1 Application to PIC Plasma Simulations

In Particle-In-Cell (PIC) plasma simulations, the space in which the plasma is to be simulated in is divided into a large number of equal sized regular cubes (called "cells"). The boundaries of any two cubes intersect along a face of the cube or not at all (here "face" includes edges and vertices). Charged "super particles" (each representing a large number of charged particles in the plasma) move freely within these cells. A two dimensional example is show in figure 7.

The particles moves freely within the cells, but the electromagnetic fields are only computed at fixed points on the cells. For instance, we might compute the electric field at each vertex in figure 7, while

Figure 7: A two-dimensional subdivision of a square into smaller squares (dashed lines), with particles (the circles) and the corresponding graph.

computing the magnetic field at the center of each edge. These fields are then interpolated to the particles' positions. Similarly, when the fields are computed, the particles' charges and velocities are moved to the fields' points in the cells that the particles are in.

To this decomposition, we can associate a graph G: the vertices are the are the cells, while the edges connect those cells which would have to communicate with one another. In figure 7, it is assumed that each cell only has to communicate with other cells it shares a face with; this might be true if the fields are computed at the center of each cell.

Many other problems in computational physics have the structure of computation on graphs. This problem, however, was the motivation for this work. Also, remapping (load balancing) is especially important for PIC plasma simulations since as the super particles move around the space being simulated, different amounts cells can have wildly different numbers of particles and thus different amounts of work.

It is worth pointing out that if the number of particles is fixed, the total work in the system is as well, and we are in the situation that the flow algorithm can run without any global communications. If particles enter and leave the system, but only slowly, the situation is almost as good.

5.2 Application to Time-shared Parallel Processors

Many modern parallel computers are used in a time-sharing mode, so multiple jobs may be running on the same processor, while at the same time a job will be running on multiple (shared) processors. This can increase access to the machine as well as use CPU cycles that would otherwise be wasted waiting for I/O or messages to arrive. To a job running on several processors, it will appear that the job is running on a heterogeneous parallel machine with changing processor speeds.

A similar situation occurs when a cluster of workstations (on a LAN) is used as a low-end parallel processor. In addition to possibly other parallel jobs, processors may be used by an interactive user. In addition to speeding up a parallel job, reducing the load on a machine is necessary if the interactive user is not going to get annoyed and perhaps remove the machine from those available for use by parallel jobs.

In this case, often there will be only a few changes in the speed of a given processor (as a new job arrives or an interactive user logs in). If so, we will again be close to the situation in which the flow algorithm does not need global communications.

6 Results

Unfortunately, I do not have any results to present yet beyond those given in [21]. While a good deal of work was accomplished on the test bed program over this summer, it was only for the last few weeks that

I was able to concentrate entirely on this project. (Before that, I prepared a talk for the SIAM Discrete Mathematics conference, and was waiting to see if Rome would have a project of their own for me to work on.)

I expect to have the test bed program ready to do some tests in the near future.

References

[1] Stephen T. Barnard and Horst D. Simon. Fast multispectral implementation of recursive spectral bisection for partitioning unstructured problems. *Concurrency*, 6(2):101–117, 1994.

[2] P. Ciarlet, Jr. and F. Lamour. On the validity of a front-oriented approach to partitioning large sparse graphs with a connectivity constraint. preprint, 1994.

[3] X. Deng, H. Liu, J. Long, and B. Xiao. Deterministic load balancing in computer networks. In *Proceedings of the IEEE Conference on Parallel and Distributed Processing*, 1990.

[4] C. M. Fiduccia and R. M. Mattheyses. A linear-time heuristic for improving network partitions. In *19th IEEE Design Automation Conference*, pages 175–181, 1982.

[5] P. Fizzano. *Centralized and Distributed Algorithms for Network Scheduling.* PhD thesis, 1995.

[6] L. R. Ford, Jr. and D. R. Fulkerson. *Flows in Networks.* Princeton University Press, Princeton, NJ, 1962.

[7] M. R. Garey and D. S. Johnson. *Computers and Intractability.* Freeman, New York, 1979.

[8] Jonathan L. Gross and Thomas W. Tucker. *Topological Graph Theory.* Wiley Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, New York, 1987.

[9] Ajay K. Gupta and Susanne E. Hambrusch. Load balanced tree embeddings. *Parallel Computing*, 18:595–614, 1992.

[10] Alan Heirich. Scalable load balancing by diffusion. preprint, Caltech, 1994.

[11] Bruce Hendrickson and Robert Leland. An improved spectral graph partitioning algorithm for mapping parallel computations. Sand92-1460, Sandia National Laboratories, Albuquerque, NM, 1992.

[12] Bruce Hendrickson and Robert Leland. Multidimensional spectral load balancing. Sand93-0074, Sandia National Laboratories, Albuquerque, NM, 1993.

[13] Bruce Hendrickson and Robert Leland. A multilevel algorithm for partitioning graphs. Sand93-1301, Sandia National Laboratories, Albuquerque, NM, 1993.

[14] B. Hoppe and E. Tardos. The quickest transhipment problem. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1995.

[15] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. 95-035, University of Minnesota, Department of Computer Science, Minneapolis, 1995.

[16] B. W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal*, 49:291–307, 1970.

[17] Donald E. Knuth. Two notes on notation. *American Mathematical Monthly*, 99(5):403–422, May 1992.

[18] Joseph W. H. Liu. A graph partitioning algorithm by node seprators. *ACM Transactions on Mathematical Software*, 15(3):198–219, 1989.

[19] Michael E. Palmer, Marc Rieffel, Stephen Taylor, and Jerrell Watts. The concurrent graph: Basic technology for irregular problems. *IEEE Parallel and Distributed Technology*, 1995. (In Press).

[20] Alex Pothen, Horst D. Simon, and Kang-Pu Liou. Partioning sparse matrices with eigenvectors of graphs. *SIAM Journal of Matrix Analysis and Applications*, 11(3):430–452, 1990.

[21] Mark Purtill. Static and dynamic graph embedding for parallel programming. Final report for afosr summer faculty research program, Phillips Laboratory, 1994.

[22] Mark Purtill. A C++ test bed for graph partitioning and mapping for parallel programming. Final report for afosr summer faculty research program, Phillips Laboratory, 1995.

[23] John E. Savage and Markus G. Wloka. Parallelism in graph-partitioning. *Journal of Parallel and Distributed Computing*, 13:257–272, 1991.

[24] Jianjian Song. A partially asynchronous and iterative algorithm for distributed load balancing. *Parallel Computing*, 20:853–868, 1994.

[25] Peter R. Suaris and Gershon Kedem. An algorithm for quadrisection and its application to standard cell placement. *IEEE Transactions on Circuits and Systems*, 35(3):294–303, 1988.

[26] Stephen Taylor and Jerrell Watts. A practical approach to dynamic load balancing. *IEEE Transactions on Parallel and Distributed Systems*, 1995. (submitted).

[27] Marc H. Willebeek-LeMair and Anthony P. Reeves. Strategies for dynamic load balancing on highly parallel computers. *IEEE Transactions on Parallel and Distributed Systems*, 4(9):979–993, 1993.

DETECTION OF CONCEALED OBJECTS IN IMAGES: INVESTIGATION INTO WAVELET TRANSFORM BASED OBJECT ISOLATION TECHNIQUES

Mysore R. Raghuveer
Associate Professor
Department of Electrical Engineering

Rochester Institute of Technology
79 Lomb Memorial Dr.
Rochester NY 14623-5603

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base
Washington DC

and

Rome Laboratory

September 1996

DETECTION OF CONCEALED OBJECTS IN IMAGES: INVESTIGATION INTO WAVELET TRANSFORM BASED OBJECT ISOLATION TECHNIQUES

Mysore R. Raghuveer
Associate Professor
Department of Electrical Engineering
Rochester Institute of Technology

Abstract

Wavelet transform based techniques were developed and investigated for isolation and enhancement of objects in images. The primary motivation is the development of image processing algorithms as part of an automatic system for the detection of concealed weapons under a person's clothing; a problem of considerable potential utility to the military in certain common types of deployment in the post cold war environment such as small unit operations. The issue has potential for other dual use purposes such as law enforcement applications. Wavelet decompositions of the currently available database, namely, noisy, low contrast, infrared images, were studied in space-scale-amplitude space. An isolation technique for separating potential suspicious regions/objects from surrounding clutter has been proposed. Based on the images available, the study indicates that the technique is very promising in providing the image enhancement necessary for further pattern detection and classification.

DETECTION OF CONCEALED OBJECTS IN IMAGES: INVESTIGATION INTO WAVELET TRANSFORM BASED OBJECT ISOLATION TECHNIQUES

Mysore R. Raghuveer

Introduction

The problem addressed is the detection of concealed objects such as weapons underneath a person's clothing by processing digital images acquired from imaging sensors. Examples of

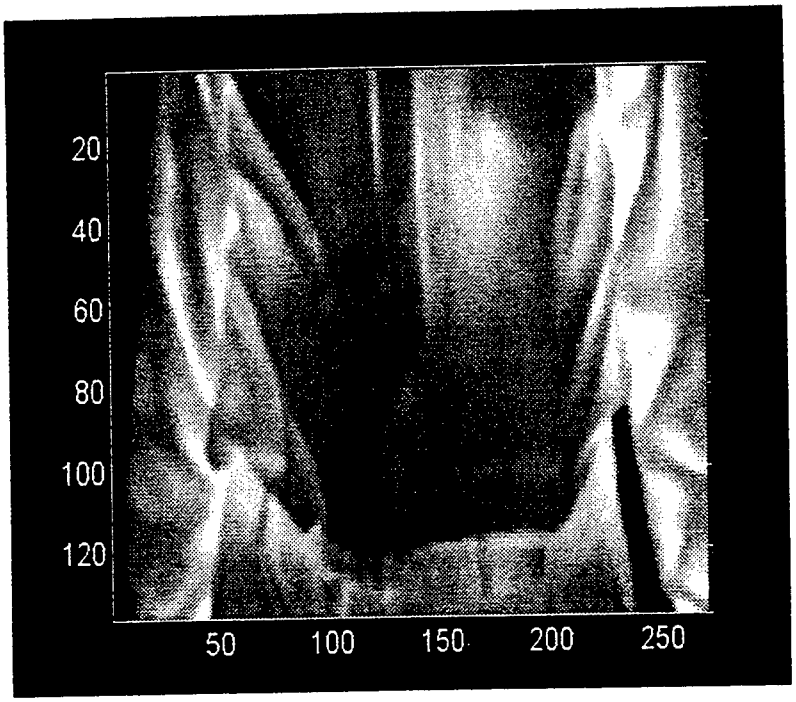


Figure 1: IR image of person with concealed handgun.

such sensors are active devices such as x-ray, radar or acoustic arrays or passive devices such as millimeter or infrared (IR) sensors. There are various situations involving monitoring for security purposes, both current and anticipated, that require the screening of people for concealed objects. One example is the overseeing of the Bosnia operations by the US military in the wake of the recent peace agreement. The task involves interaction of the US forces with large numbers of residents most of whom do not pose a threat. However, it is important to identify those who do pose a potential threat by virtue of carrying concealed weapons. An example of a dual use

application is law enforcement. In all these cases the ability to obtain images using sensors that can see the concealed objects through the clothing would be extremely valuable.

However, available imaging sensors are not ideal. Each type of sensor has its own strengths and weaknesses. The captured images are plagued by problems such as low signal to noise ratio, low contrast, poor resolution and clutter. For example, although IR sensors provide images of good resolution, they perform poorly if the target individual is wearing heavy clothing. Therefore, these applications require the development of digital image processing techniques and algorithms for (a) enhancing potential contraband objects in the image for easier viewing and

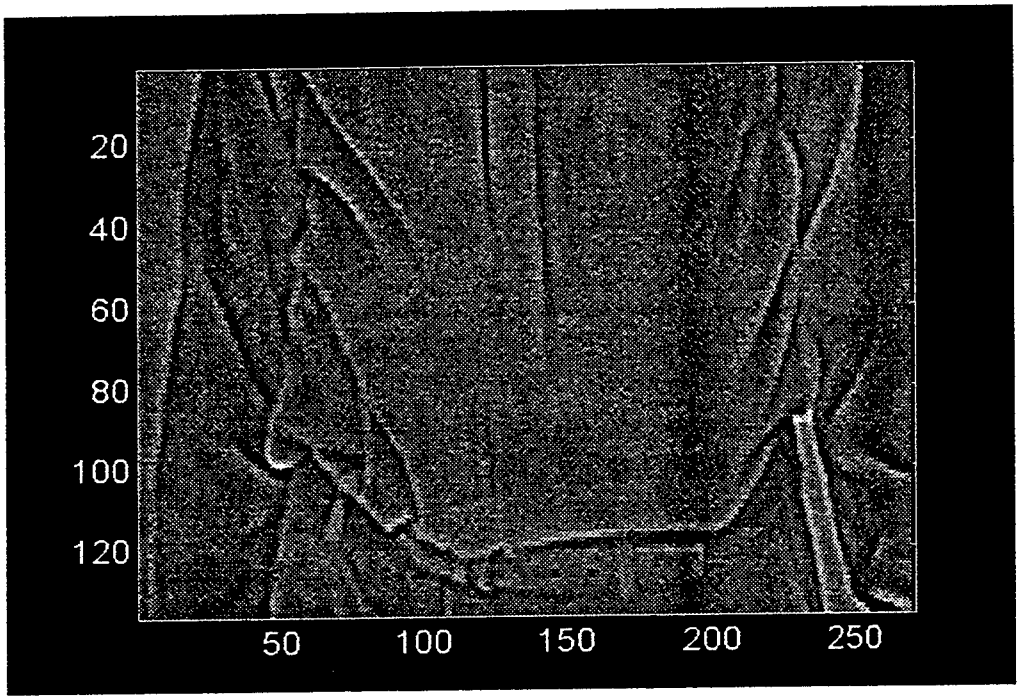


Figure 2: Laplacian of image in Figure 1

identification by trained personnel and (b) automatic detection and tagging of such objects in an image with good power of detection, that is, high probability of detection and low probability of false alarm.

Current Approaches

Even though the importance of imaging for concealed weapons detection (CWD) has been recognized as shown, for instance, by the ARPA/NIJ/ Rome laboratory program, digital

image processing efforts in this area are quite recent and there is not much previously published literature. The key publications are listed in the References section. Of these, only Felber et al [3] deal with digital processing of acquired images. They describe a CWD scheme involving radar and ultrasound. In their scheme multiple frequency radar (0.5-4 GHz) with a range capability of 10-15 m is used to monitor a crowd. The CWD capability rests on the fact that the reflection as a function of frequency varies with the target type. Thus, the contrast in reflected intensity as a function of frequency between concealed metallic and plastic weapons and other surrounding material (clothing, human skin etc.) is exploited in tagging potential threats in the crowd.

Digital imaging is performed on the ultrasound input. The ultrasound sensor kicks in on receiving a cue from the radar sensor about specific potential threats. It operates at 200 kHz and a

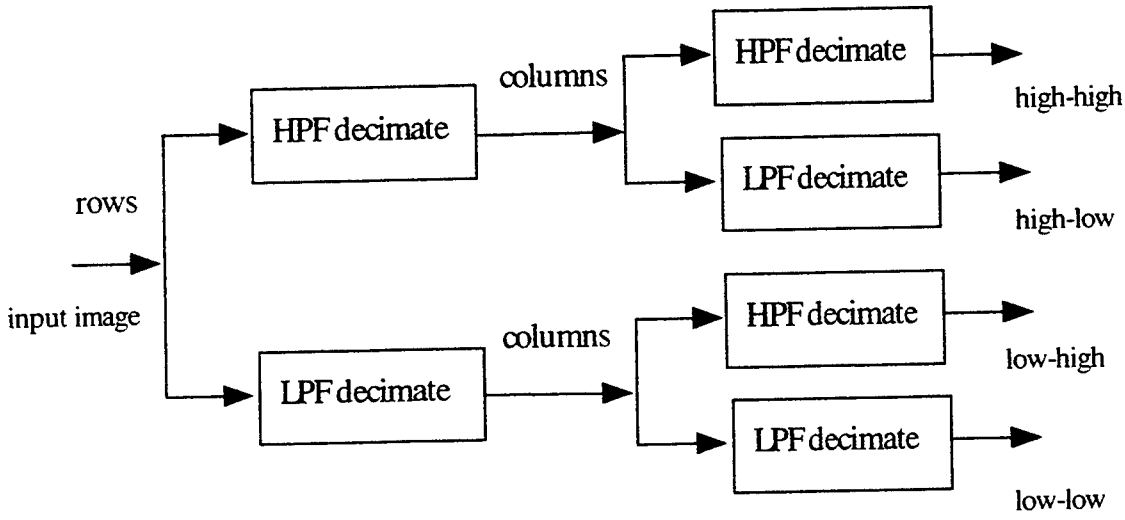


Figure 3: One stage of wavelet decomposition

range of 3-5 m. The current implementation relies on a single sensor that has to be physically moved across the subject rather than using an ultrasound sensor array. The digital image processing is done as follows. The reflected ultrasound waveform at each pixel position is first filtered to remove noise by thresholding. Next, these waveforms are Fourier transformed and limited to a band centered at the transmit frequency to filter out spurious frequencies such as, for example, power supply hum. An image is constructed using the peak value of the Fourier transform magnitude for each pixel position. A final binary thresholding is done on the image with the intent of increasing the brightness and contrast. The authors report very low false alarm

rates for this technique. Disadvantages are appropriateness for just one sensor type, long processing time and the non-exploitation of dependencies across the image.

Objectives

Right at the beginning of the summer assignment, the following objectives were set after discussions with the lab focal point:

1. Identify a problem area in CWD imaging.
2. Investigate a suitable methodology with experiments on real data.

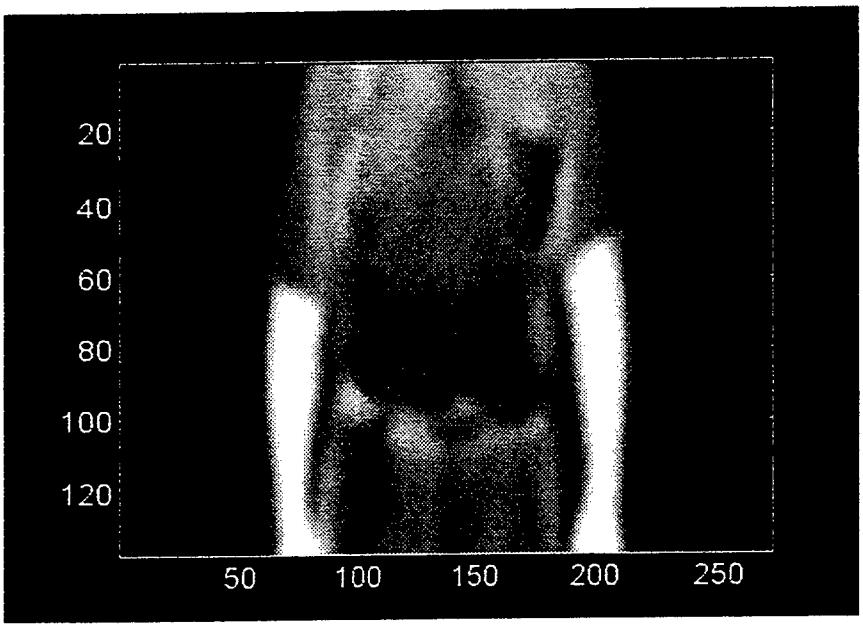


Figure 4: IR image: moderate clothing.

3. Provide results in the nature of a proof of concept.

The objectives were influenced by the shortness of the summer research period and the newness of the research area.

Identifying a Problem Area in CWD Imaging

The investigation first looked into identifying a suitable problem in imaging for CWD. Questions that initially confronted the investigator were:

- What are probably the key tasks of the image processing component of an automated CWD system?
- Can the above tasks be assigned a hierarchy in terms of complexity and dependence?
- What would be an important problem to tackle during the summer in terms of contributing towards an effective implementation of any of these tasks?

The image processing in an automated CWD system would conceivably have components at the following levels:

1. Enhancement of the acquired digital images so as to make concealed objects clearly visible and, in the process, enabling a trained person to identify threatening or contraband objects. Furthermore, the enhanced images would serve as inputs to step 2. Preferably the techniques would be sufficiently general to work with different types of imaging sensors. Simple changes in parameter inputs to the programs should be all that is necessary to account for the different sensor types. The primary impairments that the techniques should handle are noise and low contrast. Poor resolution can be a problem as well.
2. Automatic detection and classification of concealed objects in the image. In conjunction with item 1 above, automatic detection and classification would be expected to increase the power of detection. Development of these techniques requires training sets containing various concealed objects such as handguns. A possible output of the detection and classification could be pointers in an image to suspicious objects along with an associated degree of confidence.
3. Fusion of information from different sensors when two or more sensor types are used for image acquisition. This step builds on the first two.

As we move from task 1 to task 2 to task 3, the complexity increases.

The current state of the field clearly points to work required at level 1. The field is new and has some distance to go before one can look at problems in levels 2 and 3. Therefore, it was decided that addressing a problem in image enhancement would be more important than addressing a problem at the second or third level.

Methodology

The applied nature of the research and the lack of precise mathematical descriptions of the objects involved necessarily points to a predominantly heuristic investigation. The only images that were available to the investigator were those captured using infrared (IR) sensors.

However, the requirement that the developed technique be flexible to be applied to other imagery was taken into account. Infrared imaging for CWD is a passive technique. Detection of concealed objects relies on them showing up in the image as regions that are cooler than their surroundings. There are several factors that come in the way of a clear resolution of the object's signature. Obviously, the farther the distance between the IR camera and the person being imaged, the less the contrast between the object's IR signature and the surrounding. The thickness of the clothing on top of the concealed object is a determining factor of the attenuation. With heavy clothing it is virtually impossible to discern contribution from a concealed object to

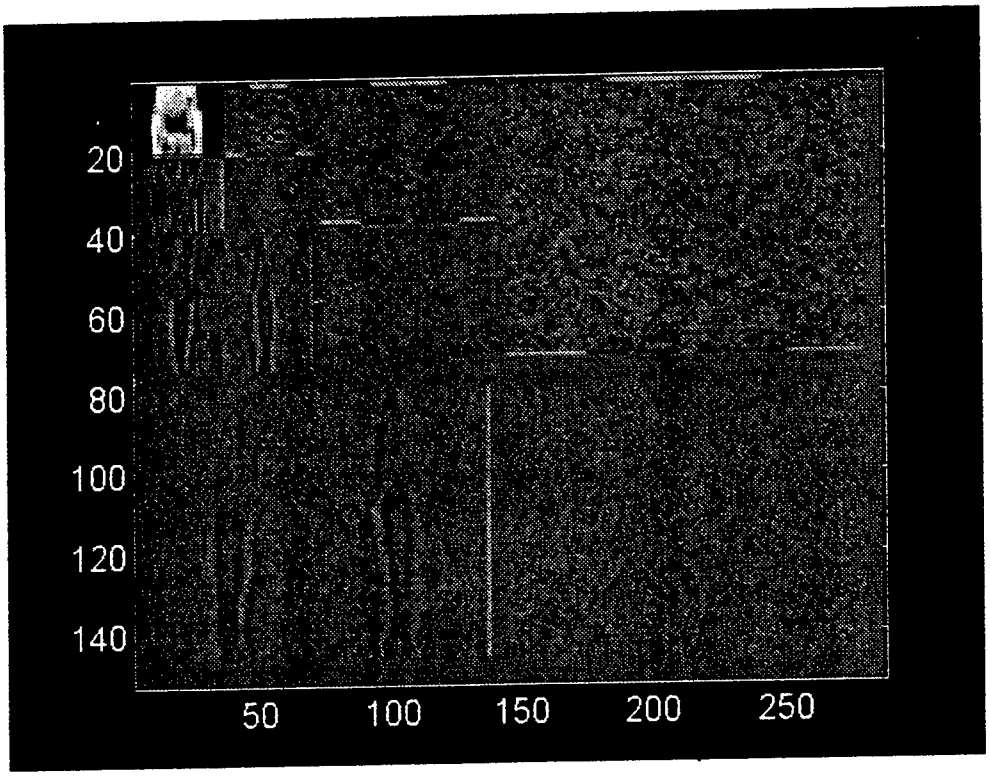


Figure 5: DWT of image in Figure 4

the image. While the factors outlined here can be regarded as contributing to low contrast and poor resolution, there are other usual impairments such as sensor noise and clutter. By clutter we mean objects in the image that are not concealed weapons. These could be items on the clothing such as buttons etc. or aspects of the scene manifesting as cool or dark regions in the image thus posing potential for false alarms.

Some of the issues mentioned in the previous paragraph are brought out in Figure 1 which shows an IR image of a person with a concealed handgun. This is an instance of light

clothing and several objects can be discerned. The gun is located near the belt. To get some idea of how discernible the handgun is in the image, a random selection of people was asked to guess what was in the image. The purpose was not to develop a rigorous score but to get a rough idea. The test indicated that once told there was a weapon most observers could tell the outlines of the handgun. However, it is difficult to gather the orientation of the weapon visually. Furthermore, there are dark patches at other locations where there are no concealed objects. The challenge then for any image enhancement algorithm is to bring out objects in greater relief so that it results in better recognition.

As has generally been recognized, edge and texture information are key inputs to many

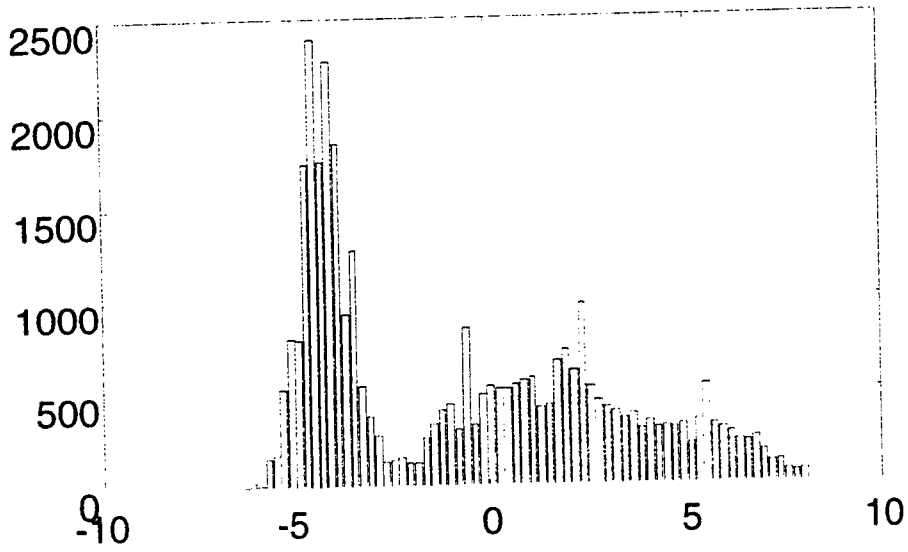


Figure 6: Histogram of original

an object recognition algorithm. For the images dealt with here there is not much of a textural signature from the objects of interest. The images as can be seen from the example are of low contrast. However, some type of edge extraction would be helpful. The problem is low contrast and diffused edges. Consequently, traditional edge enhancement methods can be expected to fail as shown by the Laplacian in Figure 2. While the hard edges are clearly visible, the weapon is not. The characteristics of the images suggest looking beyond such conventional techniques.

. The investigator's earlier research in wavelet analysis suggests the use of wavelet based techniques. The natural denoising and zooming properties of wavelet transforms can be used to

obtain scale-invariant enhancement, detection and classification [5,6]. A basis for the use of the wavelet transform in edge detection and edge mapping was presented by Mallat and Hwang [7]. Fundamentally, the technique involves classifying as edges those features of the signal or image that show as edges at all levels of the wavelet transform. However, the diffused nature of the objects and the softness of the edges in the given images point away from making such an approach the basis for enhancement. Instead, a technique aimed at object isolation was developed.

The Object Isolation Algorithm

Object isolation here is defined as the tagging of pixels in the image as belonging to a specific object on the basis of some criteria. It amounts to grouping pixels into several disjoint sets. The first step in the technique developed is discrete wavelet transformation (DWT) of the images. The intent is to achieve some amount of object isolation on the basis of the sensitivity of the wavelet transform to scale or size of an object or parts of an object. The block diagram of the transformation is shown in Figure 3. This report does not deal with the minute details of wavelet transformation and reconstruction which can be found elsewhere [8-10]. Briefly, what is involved in the first stage of the DWT is 1-D filtering and decimation by factor of two of the rows of the image by two separate one dimensional filters, one a low pass filter (LPF) and the other a (HPF). This results in two images each with half as many columns as the original. The columns of each of these images are in turn subject to the same filtering and downsampling leaving us finally with four images each with a size a quarter of the original. The "low-low" image in Figure 3 corresponds to the output of two low pass stages and is a low resolution approximation to the original. The other three images are referred to as the "detail images." A multiple level DWT is accomplished by doing the row-column filtering as above on the "low-low" image of each level. Thus an n level decomposition consists of three detail images at each level and one low resolution approximation at level n yielding a total of $3n+1$ images. The detail images contain information about edges in the image.

The notion that the DWT provides an element of scale-dependent object isolation for the test set was verified experimentally. Consider the IR image in Figure 4. The original scene is of a person with a concealed weapon under moderately heavy clothing. A three level DWT of the mean subtracted input image was obtained and is shown in Figure 5. The LPF is the Daubechies

4 tap FIR filter with coefficients *0.3415, 0.5915, 0.1585, -0.0915*. The corresponding HPF is a 4 tap FIR filter with coefficients *-0.0915, -0.1585, 0.5915, -0.3415*. Figure 5 shows that the noise and several objects that clutter the scene such as the subject's limbs, name tag etc. tend to be isolated in the outer or first two levels of detail. Consequently, several other objects stand highlighted in the low resolution approximation. That is, there is increased contrast in the residual after stripping away noise and some other objects from the image. The DWT here substantiates the assertion made earlier that the objects of interest do not show hard edges. The

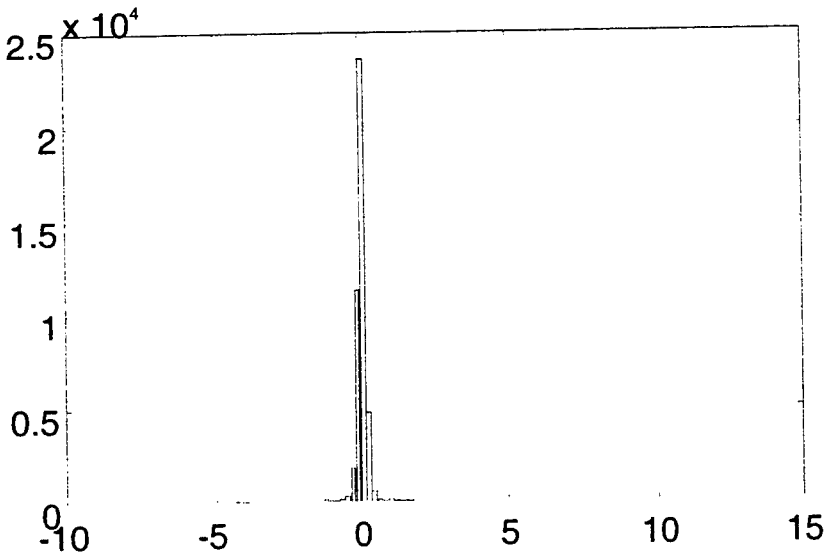
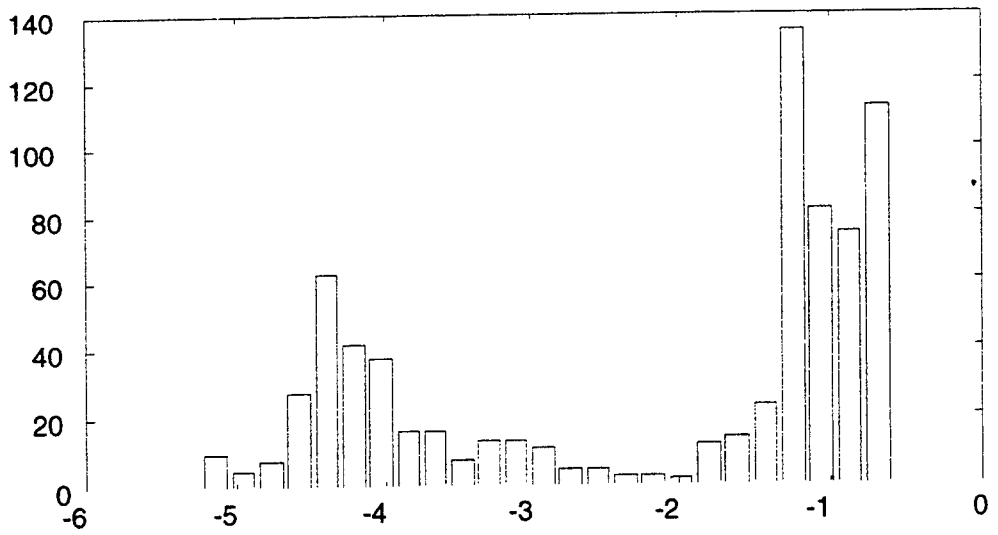


Figure 7: Histogram of DWT

concealed weapon itself does not show up as such in the DWT directly.

The next step in the algorithm is scale filtering. The outer detail functions are zeroed out since they contain noise and features not related to the weapon. The third step is a finer level of object isolation performed on what remains after scale filtering. A clue to a potentially effective approach is provided by the histogram of the DWT. Figure 6 shows the histogram of the original image while Figure 7 shows that of the DWT. Figure 6 suggests that there are two distributions in the original image and a separation can be effected with a threshold set at -2 which is a trough in the histogram. Figure 7 has a substantially different characteristic from that of Figure 6 owing to the fact that the bulk of the DWT pixels have values close to zero. Examining the histogram of the DWT at values away from the origin as in Figure 8 shows several possibilities for thresholds

Figure 8: Histogram of DWT from -5 to -0.6



at the valleys in the histogram to isolate individual distributions. One such valley or trough is seen at *-2.1*. There is a similar valley on the positive side (figure not shown) at *0.5*. Thresholding the DWT at these values and reconstructing the image results in the image shown in Figure 9 which clearly brings out the imprint of the concealed handgun. Reconstruction with alternative thresholds isolated other features. The enhancement as provided in Figure 9 makes subsequent feature extraction easier. For example, doing an edge extraction by slicing (equivalent to a contour plot) the image in Figure 9 and superimposing it on the original results in the image of Figure 10. A n examination of the traces reveals the one pointing to the concealed weapon. This is the trace that resembles letter 'L' tilted at 60 degrees.

The image in Figure 1 suffers from fewer impairments than the one in Figure 4. Figures 11 and 12 show the isolated weapon and the related edge identified images respectively. Figure 13 shows an original image taken under very noisy conditions and would be regarded as a "tough" image. The corresponding object isolated and edge detected images are shown in Figures 14 and 15. The concealed weapon is contained within the elongated contour in the center. This image also reveals edges that could be interpreted as concealed weapons even though they are not.

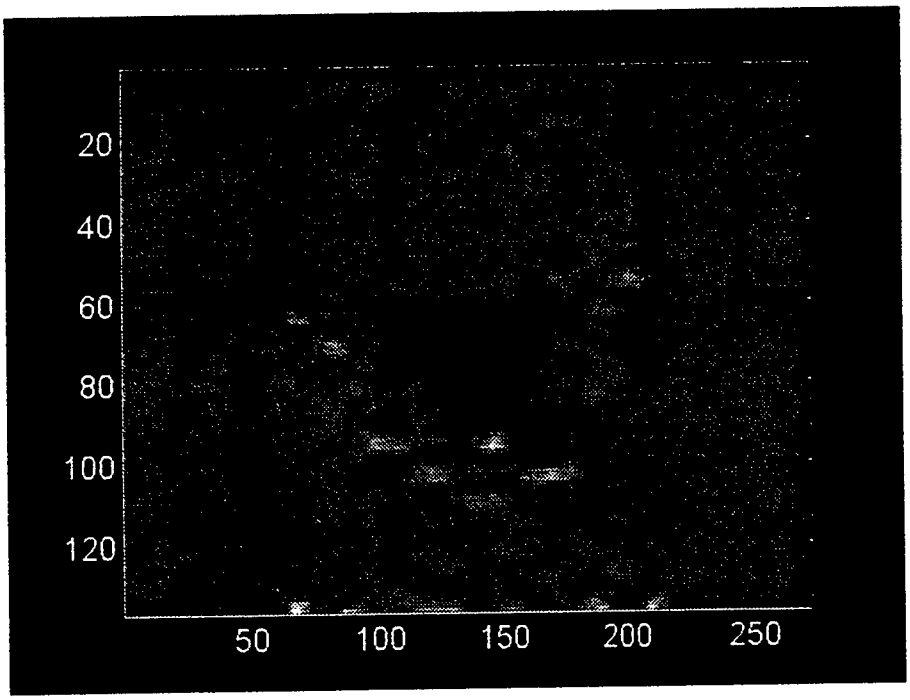


Figure 9: Result of object isolation.

The steps of the algorithm are summarized below:

1. Find the DWT of the mean subtracted image. At this time, the Daubechies 4 tap filters are being used for simplicity.
2. Generate a histogram of the DWT.
3. Scale filter the DWT.
4. Use valleys or troughs of DWT histogram to set thresholds. Generate a reconstruction for every interval defined by two successive thresholds. This involves retaining values of the DWT within this interval and zeroing out the rest.
5. Generate an edge map for each reconstruction. For want of space, only the reconstruction that holds the concealed weapon has been presented for each image considered.

Figure 10: Edge map superimposed on original

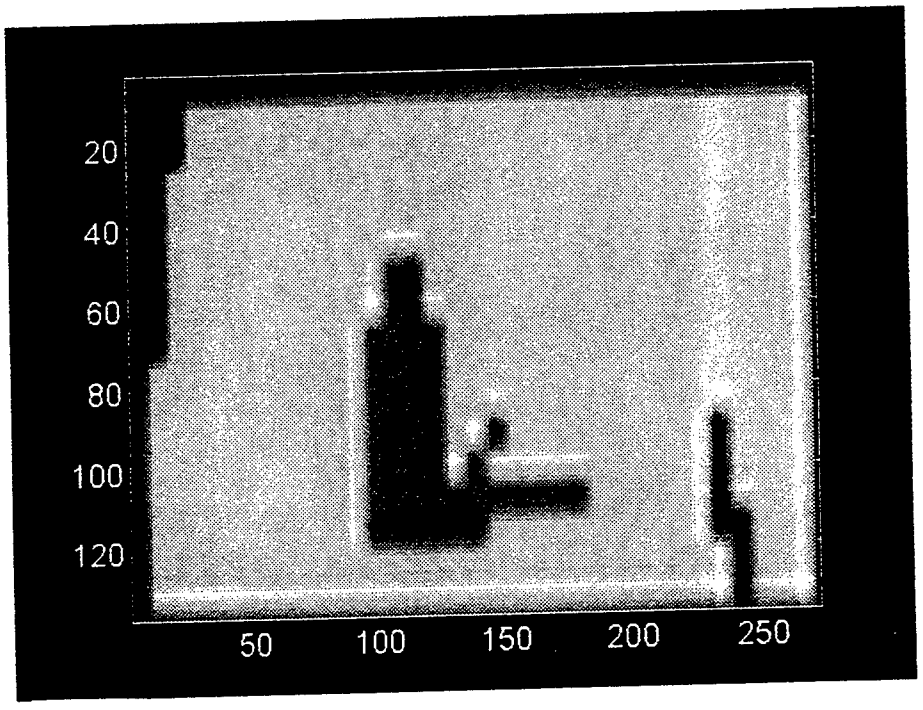


Figure 11 Isolated imprint of concealed weapon for image of Figure 1.

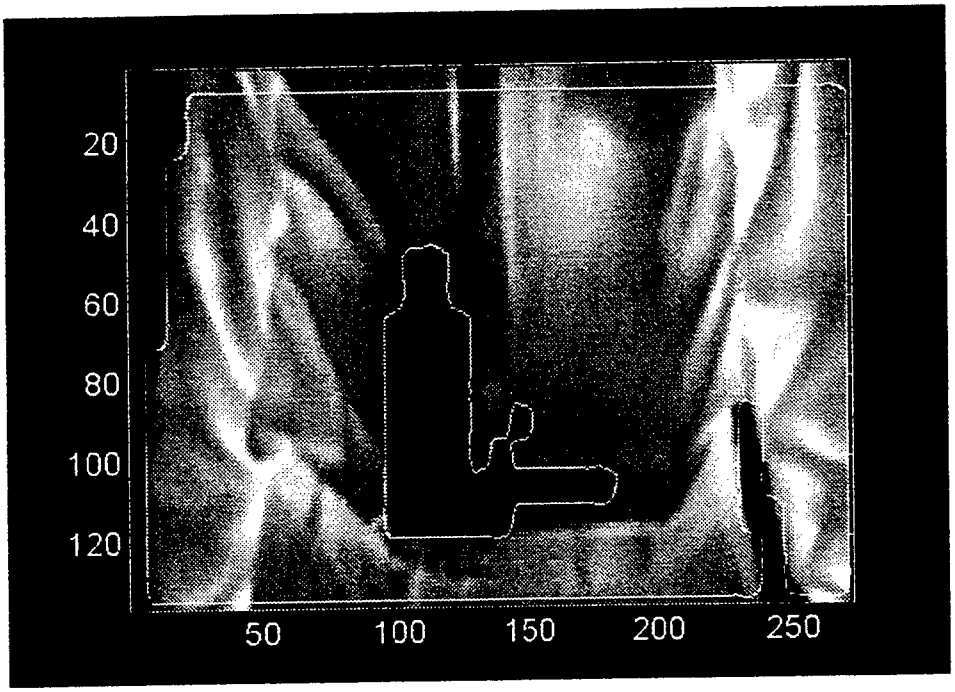


Figure 12: Edge map

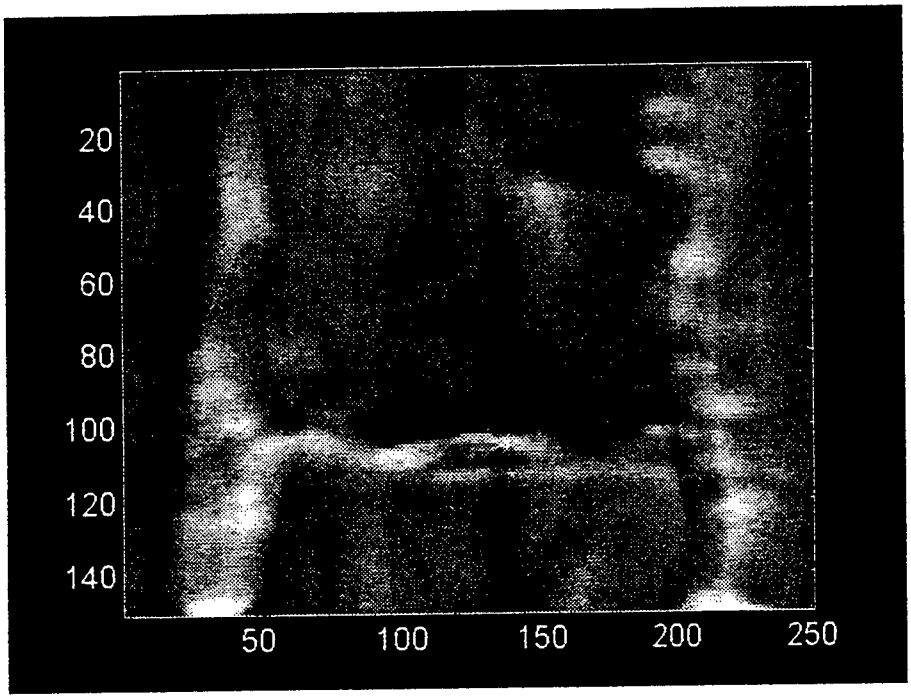


Figure 13: Image#3

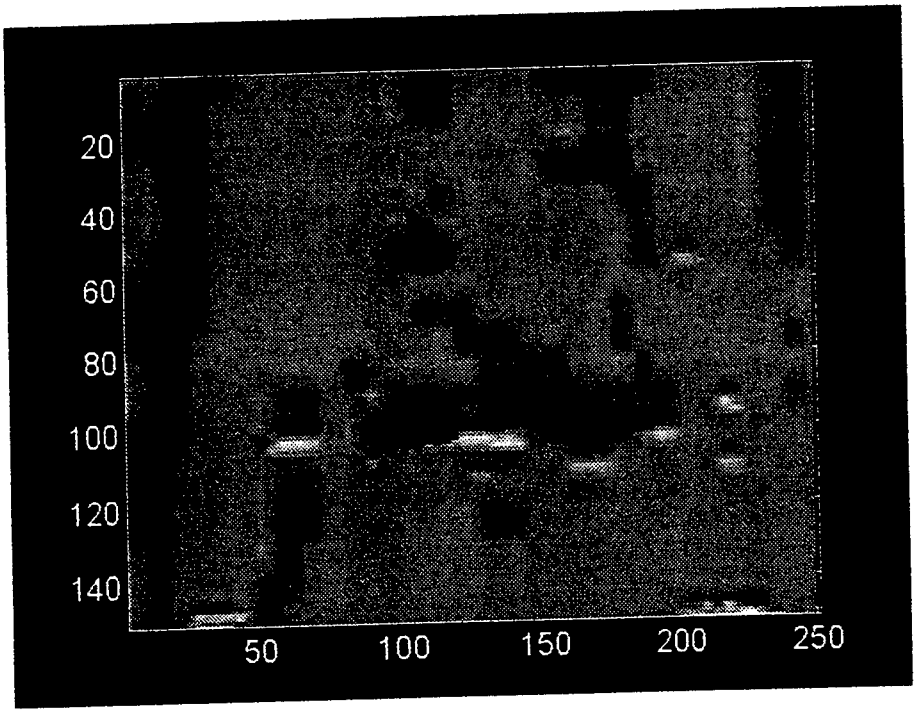


Figure 14: Result of object isolation

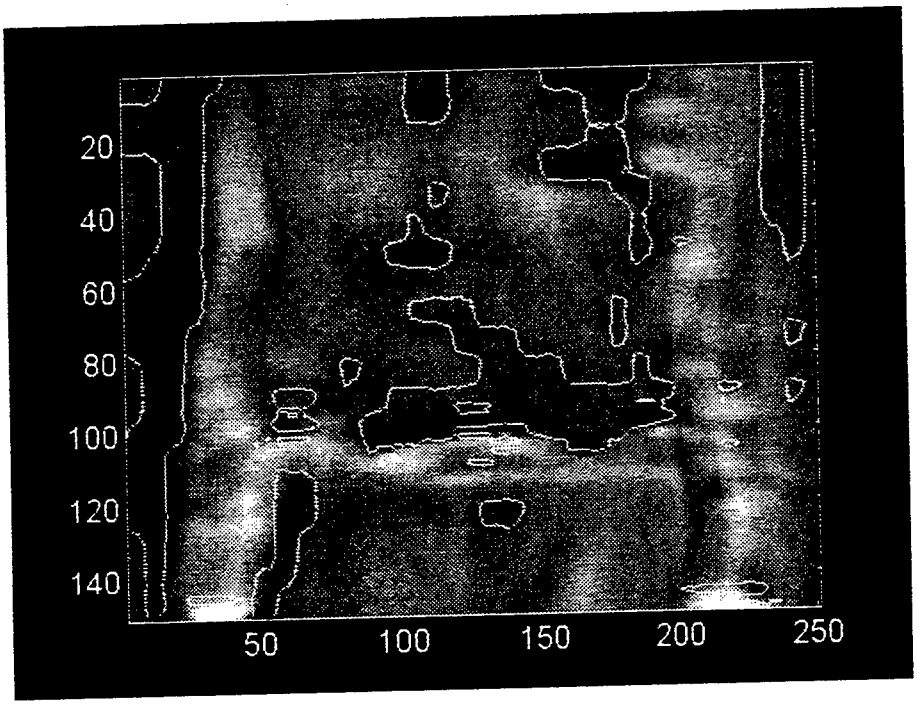


Figure 15: Edge map

Summary and Future Work

The objectives set at the beginning of the summer assignment were fulfilled. The wavelet based technique has been demonstrated to hold potential for use in a more complex automatic CWD system by virtue of its ability to provide object isolation. Although the experimental data was of one sensor type: infrared, the imaging technique is applicable to other digital imagery. The technique is new in that it focuses on statistical grouping of DWT data based on its histogram rather than the edge extraction approaches used by most investigators in other object recognition problems.

Further testing and refinement of the algorithm requires a larger test set with multiple sensors. Among the things to be characterized are:

- Noise sensitivity of the technique.
- Development of a detection scheme that uses the object isolated image.
- The power of detection, that is, probability of detection and false alarm rates of the detection method.

The scheme developed is applicable to object isolation for other types of images such as medical images where potentially malignant objects have to be isolated and characterized.

Acknowledgment: The investigator is pleased to acknowledge the enormous help he received from his lab focal point Dr. James Michels, Dr. Adel Slamani, and Mr. David Ferris.

References

1. Currie, N.C., Demma, F.J, Ferris, D.D., McMillan, R.W. and Wicks, M.C., "ARPA/NIJ/Rome Laboratory concealed weapon detection program: an overview," in *Signal Processing, Sensor Fusion, and Target Recognition V,* SPIE Proc., vol. 2755, eds: I. Kadar and V. Libby, pp.492-502, Orlando, April 1996.

2. Currie, N.C., Ferris, D.D., Demma, F.J, McMillan, R.W. and Wicks, M.C., "Concealed weapon detection systems," *Rome Laboratory Technical Journal*, vol. 2, no. 1, 1996.

3. Felber, F.S., Davis, H.T., Mallon, C. and Wild, N.C., "Fusion of radar and ultrasound sensors for concealed weapons detection," in *Signal Processing, Sensor Fusion, and Target Recognition V,* SPIE Proc., vol. 2755, eds: I. Kadar and V. Libby, pp.514-521, Orlando, April 1996.

4. Sheen, D.M, McMakin, D.L., Collins, H.D., Hall, T.E., and Severtsen, R.H, "Concealed explosive detection on personnel using a wideband holographic millimeter-wave imaging system," in *Signal Processing, Sensor Fusion, and Target Recognition V,* SPIE Proc., vol. 2755, eds: I. Kadar and V. Libby, pp.503-513, Orlando, April 1996.

5. Chapa, J.O., *Matched wavelet construction and its application to target detection.* Ph.D. dissertation. Rochester Institute of Technology. August 1995.

6. Raghuveer, M. R. and Chapa, J. O., "Object detection through matched wavelet transforms" (Invited Paper), *Wavelet applications III,* SPIE Vol. [2762-03], Ed.: H. Szu, pp. 45-50, Orlando, FL, April 1996.

7. Mallat, S. and Hwang, W. L., "Singularity detection and processing with wavelets," *IEEE Transactions on Information Theory,* v. 38, pp. 617-43, March 1992.

8. Mallat, S. "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp.674-693, July 1989.

9. Kaiser, G., *A friendly guide to wavelets.* Birkhauser, Boston, MA, 1994.

10. M. Vetterli and J. Kovacevic, *Wavelets and subband coding.* Prentice-Hall, 1995.

Integrating a Multimedia Database and WWW Indexing Tools

Scott Spetka
Associate Professor
Department of Computer Science

State University of New York
Institute of Technology at Utica/Rome
Route 12 North
Utica, New York 13504

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

September 1996

Integrating a Multimedia Database and WWW Indexing Tools

Scott Spetka
Assistant Professor
Department of Computer Science
State University of New York
Institute of Technology at Utica/Rome

Abstract

The dropping cost of computing and data communication has lead to rapidly growing data networks. Production systems are evolving to include access to on-line multimedia information while powerful workstations replace character terminals. Exploiting the modern environment requires easy user-level access and an architecture that facilitates data integration. This paper describes an architecture which features WWW indexing tools that extend the functionality of distributed relational databases. It provides access through standard HTML-based WWW browsing tools to multiple databases. The resulting distributed system is capable of growth and is accessible world-wide.

Integrating a Multimedia Database and WWW Indexing Tools

Scott Spetka

1 - Introduction

Integrating databases and providing access through wide-area networks increases the value of the data significantly while reducing maintenance costs. Shared access reduces the need for redundant copies of data and associated update problems that may affect system reliability. The Web Data Server (WDS), developed at Rome Laboratory, addresses these problems by providing general purpose access to databases, based on World-Wide Web (WWW) technology. This approach allows connection from any location on a wide-area network to databases distributed throughout the network. The system depends on user access to standard tools for multimedia data access and display that are widely distributed and in most cases available free of charge. Anyone with appropriate authorization can access the WDS system.

This paper presents a case study of the WDS system development project. Except for commercial databases which it is designed to access, the system has evolved from freeware collected through the Internet over the past year. This paper describes the system architecture and the design of the interface between existing multimedia database access functions and WWW indexing technology. System performance issues are also discussed along with consideration of implementation alternatives.

2 - The WDS Architecture

The WDS architecture is designed to provide the interface between existing multimedia database systems

and WWW technology. This section describes the evolution of the system and design issues faced by the system architects. It describes current work and future plans for extending the system.

2.1 - WDS Release 1

The WDS server is built upon an existing image database application program interface (API). The initial release of the system (see figure 1) demonstrated access to the image database through cgi-bin extensions to the WWW server. The WDS accepts commands from standard WWW browsers and responds to the commands by formatting data for output through the browsers. WWW browsers arrange for execution of appropriate tools when objects returned to tbe browser cannot be handled by built-in functions. Cgi-bin functions retrieve products from the database through the image database API as illustrated in the figure.

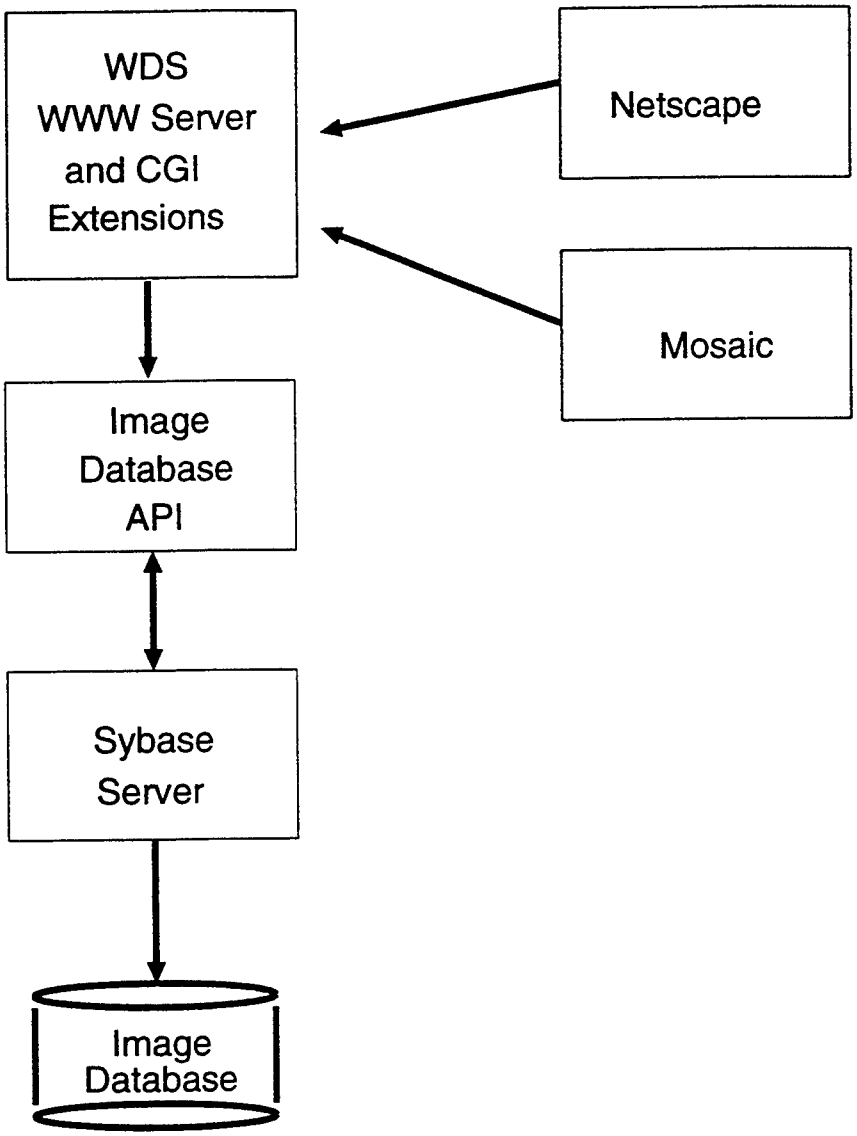


Figure 1: WDS Architecture - Release 1 - May 1995

2.2 - WDS Release 2

Figure 2 shows the release 2 WDS architecture. The Harvest indexing system provides convenient access to

the image database through standard WWW tools. The data that is used to build indexes and to support the browse capability is maintained by the WDS openserver. The openserver provides the principal interface between WWW technology and existing image databases is described below. The Release 2 architecture is shown in figure 2 below.

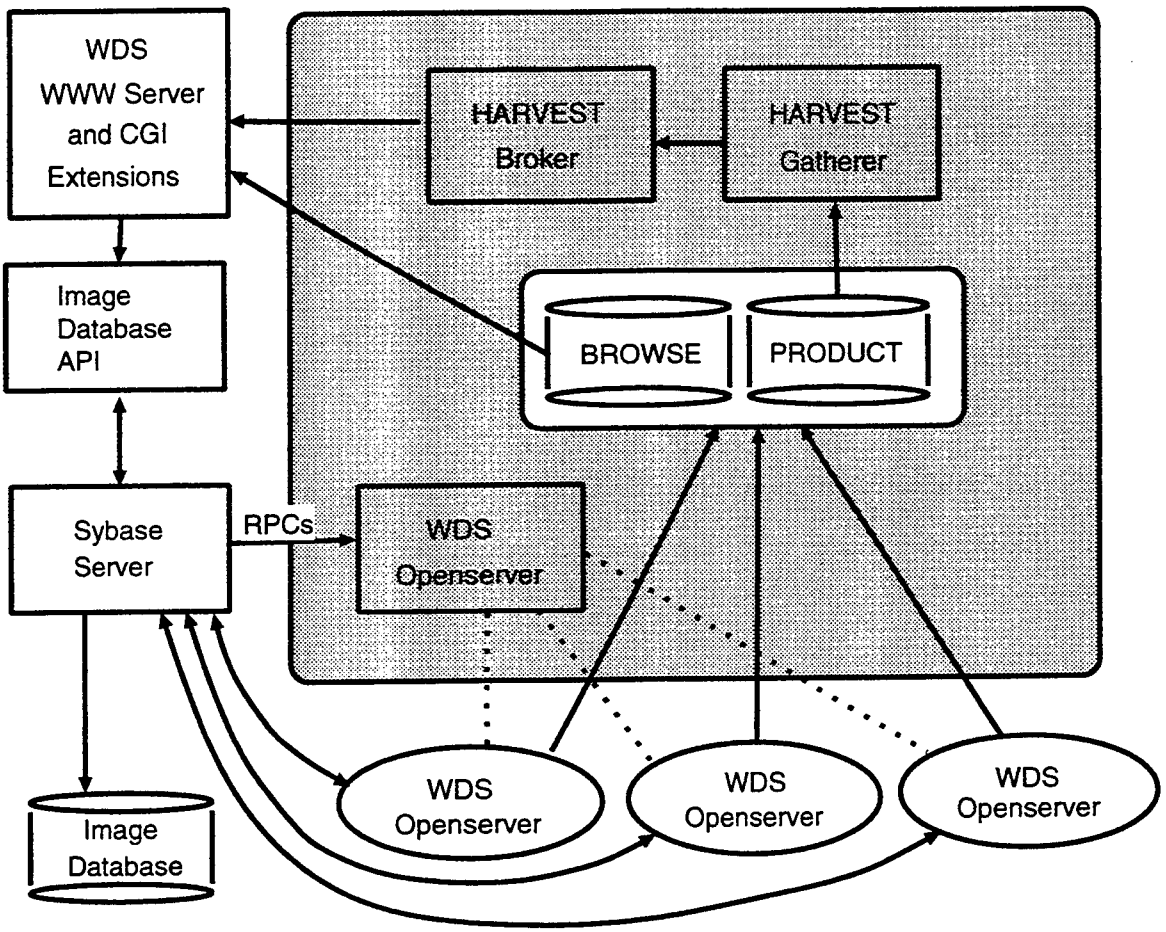


Figure 2: WDS Architecture - Release 2 - May 1996

The HARVEST Index

WDS supports two distinct interfaces for indexed product access, direct access and browse. Each interface depends on a HARVEST index through all image keywords or through attributes selected from the image database for browsing. Keeping these indexes up-to-date is done in two ways. A utility program is run to build indexes from the current image database. The utility program communicates directly with the Sybase Server. The indexes can be rebuilt at any time through this relatively expensive process, for example after recovery from a crash. In addition, the system supports dynamically extending the indexes when products are inserted into the database. The WDS Openserver, described below, provides data to the Harvest Broker/Gatherer system (the PRODUCT and BROWSE files shown in figure 2) that is used by that system to incrementally update the HARVEST index.

The WDS Openserver

The WDS Openserver implements database triggers for each multimedia object inserted into the image database. The data inserted is passed to the WDS Openserver which is responsible for coordinating the transfer of the data required for indexing to the Harvest system. In the release 2 implementation, required data is copied into files which are periodically processed by Harvest and incorporated into the HARVEST indexing system.

The Harvest system is currently invoked through a "cron" timer to incrementally update the browsing indexes for the HARVEST server. While the Harvest system is running, updates to the database are blocked to assure consistency. This functionality may be moved into the WDS Openserver so that incremental update can be performed at time intervals or after a threshold of product update activity is exceeded.

WDS supports a browsing mode where users are presented with a list of keywords that can be found in the database. The most important design constraint for the WDS openserver was to assure that data found in the BROWSE mode would produce query results from the image database. Keywords can only be inserted into the browse list after being indexed by the Harvest system. Similarly, if an image is deleted from the database it must immediately disappear from the browse list. Without this constraint it would be possible to select a keyword from the browse list only to find that there were no matching products.

2.3 - Multi-Database Support

The WDS architecture was designed to provide access to distributed multimedia database systems in a wide-area network. In addition to designing the user interface to allow selection of multiple databases, the WWW interface was designed to support multiple databases as well. The Harvest system, used for gathering data and building an index, supports access to data from multiple sources. Multiple image databases are supported without changing cgi-bin functions, so long as all databases accessed support the generic image database API. The WDS server can easily accommodate the index update functions for additional Sybase databases as illustrated in figure 3 below. A separate WDS server is configured to respond to updates for each of the target databases. Accommodating additional database servers, such as Oracle, should be relatively easy so long as they support a client/server execution model. Figure 3 shows a multi-database configuration of WDS.

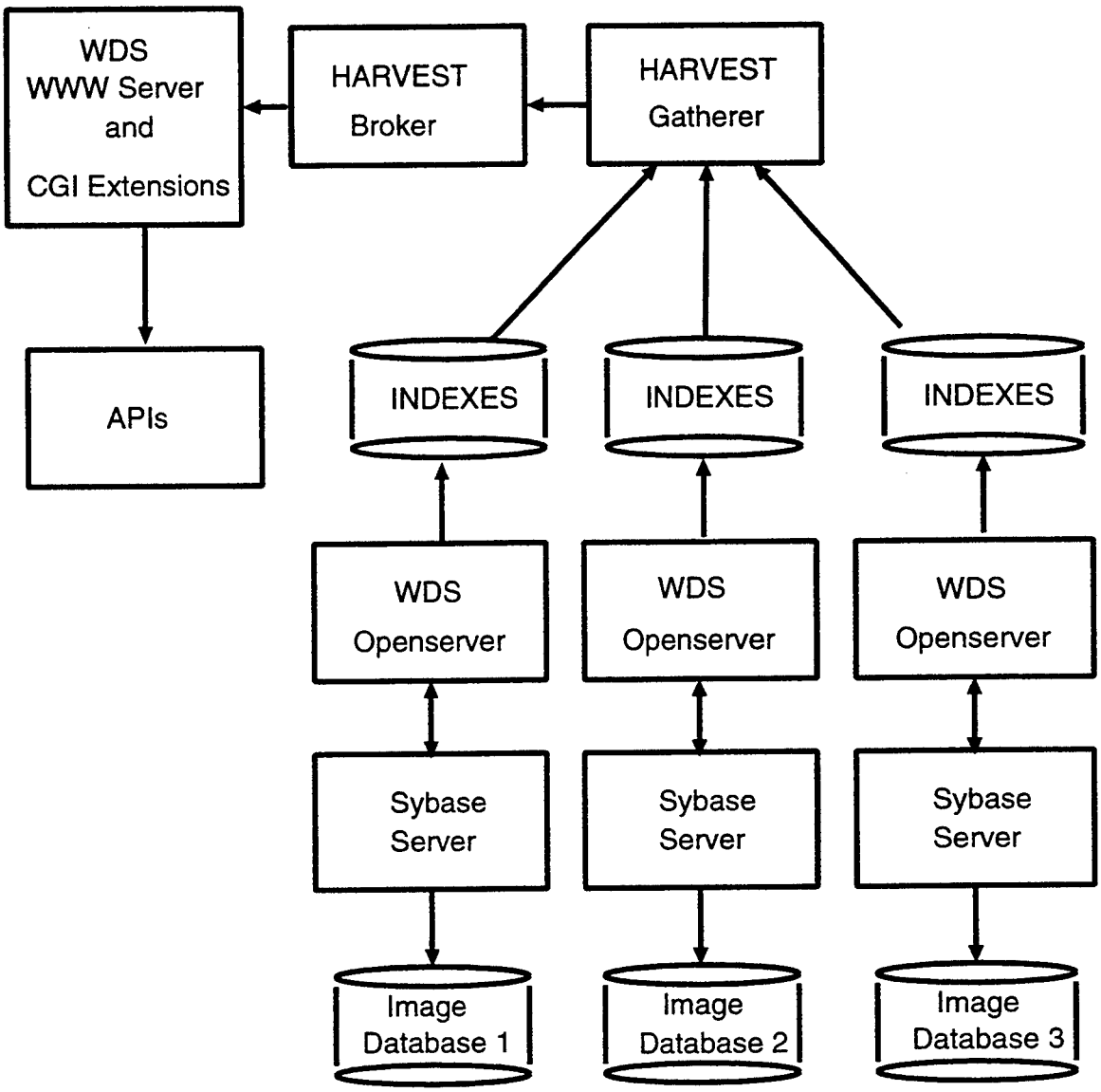


Figure 3: WDS Multi-Database Architecture

2.4 - WDS Performance

The first step in evaluating system performance is to detect possible bottlenecks in the WDS system to focus the study. Figure 1 illustrates the architecture of the WDS interface.

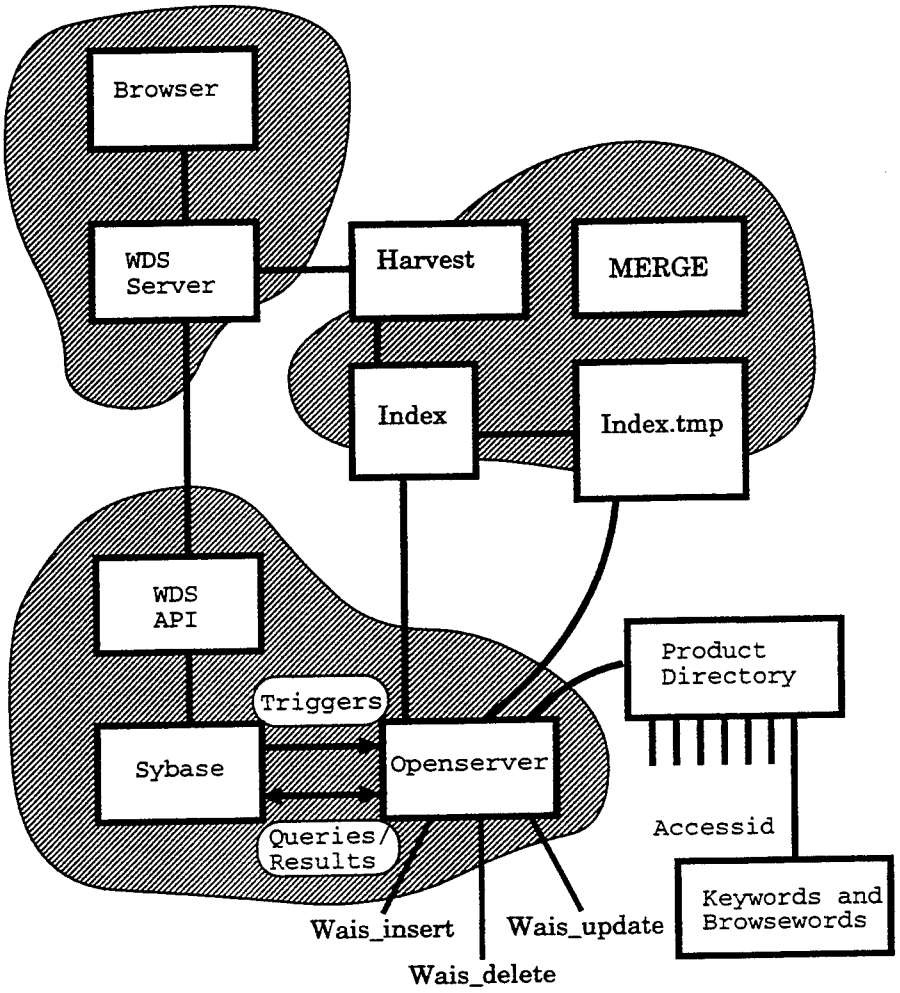


Figure 4: WDS Performance Bottlenecks

Measuring Performance

The WDS Server/Browser interface can be exercised by an automated process that generates queries and sends them to the server. The load on the server can be increased to determine, from the end-user's perspective, the expected response time under a heavy load. The freeware WWWperf or a TCL/Expect script may be used to generate the load.

The ability of the openserver to handle update requests will also be examined. This experiment will help to determine the expected maximum throughput for a system that is being heavily updated. The WWWperf and TCL/Expect tools can be used for this experiment as well as to set up experiments that measure the system response under mixed loads.

The other potential bottleneck shown in the figure 4 is the process that updates the *Harvest* index. Measuring this process will allow the update interval to be fine-tuned to minimize processing time while providing maximum availability. The Harvest index must be incrementally updated when the database is modified.

Implementation Tradeoffs

Detecting performance bottlenecks naturally leads to reevaluation of implementation techniques. The WDS interface uses file-based indexing techniques to avoid dependence on specific database indexing capabilities and to provide input for WDS browse functions. The WDS system queries the underlying multimedia database, independent of the low-level IPL database interface. The independence of the WDS interface of any specific database system provides maximum portability between interfaces. Porting to a new interface requires only that the subsystem that builds a WWW index from the database be modified to access the new database or to be adapted to the content of additional interfaces.

An alternative to the WDS indexing system could be built by adding secondary indexes to all tables containing keywords used for product access by the system. Although this approach would guarantee indexed access to every element in the database through WDS, it would impose a significant performance penalty during normal operation. By using simple triggers on updated database records, the WDS system waits and processes updates in a batch. Locking techniques are implemented to assure the integrity of the database while allowing the incremental index update procedure to proceed in parallel with continued querying activity on the database.

3 - Conclusion

This paper describes an approach to integrating WWW technology with a commercial multimedia database. The approach is put into perspective by describing the evolution of the system from it's early stages to support for a distributed multidatabase architecture. Continued development of the system will test whether it conveniently supports integration with other multimedia databases in a more heterogeneous environment. A performance study will determine the degree to which suspected data access and processing bottlenecks affect system performance.

Author Biography

Scott Spetka received his Ph.D. degree in computer science from UCLA in 1989. He is currently an Associate Professor in the Computer Science Department at the State University of New York Institute of Technology at Utica/Rome. His research interests are in the areas of distributed databases, operating systems and networks. Scott has developed a network of PCs running the Unix operating system. Before becoming involved in WWW research, Scott was developing SUNY Nodes, a Network-Oriented Data Engineering System. The system is used to experiment with query processing techniques.

References

[Spetka 96] Spetka, S.E., "Network Design, Configuration and Management Issues in an Academic Institution", 1996 IEEE Dual-Use Technology & Applications Conference, ON Center, Syracuse, NY, June 1996.

[Salerno 96] Salerno, J. ,Spetka, S.E., Mozloom, P., Miller, R., Peck, D., "Intelink: Using World-Wide Web Technology for Integrating Distributed Databases", 1996 IEEE Dual-Use Technology & Applications Conference, ON Center, Syracuse, NY, June 1996.

[Spetka 95] Spetka, S.E., "Using the TkWWW Robot to Integrate Database Systems and Internet Technology", 1995 IEEE Dual-Use Technology & Applications Conference, SUNY Institute of Technology, May 1995.

[Spetka 95] Spetka, S.E., "Cost-Effective Distributed Computing", 1995 IEEE Dual-Use Technology & Applications Conference, SUNY Institute of Technology, May 1995.

[Miller 95] Miller, R. and Spetka, S.E., "Tools for Internet Collaboration: A Survey", 1995 Dual-Use Technology & Applications Conference, SUNY Institute of Technology, May 1995.

[Spetka 94] Spetka, S.E., "The TkWWW Robot: Beyond Browsing", Proceedings of the Second International WWW Conference 1994, Mosaic and the Web, Chicago, October, 1994.

Confined Optical Phonon Modes in Si/ZnS Superlattices

Gang Sun
Assistant Professor
Engineering Program/Physics Department

University of Massachusetts at Boston
100 Morrissey Blvd.
Boston, MA 02125

Final Report for:
Summer Faculty Research Program
Rome Laboratory
Hanscom Air Force Base

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

September 1996

Confined Optical Phonon Modes in Si/ZnS Superlattices

Gang Sun
Assistant Professor
Engineering Program/Physics Department
University of Massachusetts at Boston

ABSTRACT

The confinement of optical modes of vibrations in a superlattice consisting of polar and nonpolar materials is described by a continuum model. Specifically, the structure under investigation is the Si/ZnS superlattice. Optical phonon modes in Si and ZnS layers are totally confined within their respective layers since both layers can be treated as infinitely rigid with respect to the other layer. Since there are no associated electric fields with nonpolar optical phonons in Si layers, only mechanical boundary condition needs to be satisfied for these nonpolar optical modes at the Si-ZnS interface. The optical phonons in Si layers can be described by guided modes consisting of an uncoupled s-TO mode and a hybrid of LO and p-TO modes with no interface modes. In ZnS layers, a continuum model hybridizing the LO, TO and IP modes is necessary to satisfy both the mechanical and electrostatic boundary condition at the heterointerface. A numerical procedure is provided to determine the common frequency between LO, TO, and IP modes. Analytical expressions are obtained for the ionic displacement and associated electric field as well as scalar and vector potentials. These expressions can be employed directly in calculating the carrier interaction with optical phonons in the superlattice.

Confined Optical Phonon Modes in Si/ZnS Superlattices

Gang Sun

I. INTRODUCTION

With the demonstration of the InGaAs/AlInAs intersubband quantum cascade laser at $\lambda = 4.2\mu m$[1, 2], there has been interest in the possible utilization of silicon as the optically active material because of its integrability in advanced silicon microelectronics[3, 4]. In addition, there is interest in moving the lasing from the far and midinfrared range to the near infrared optical communication wavelengths, $\lambda = 1.3 \sim 1.55\mu m$[5]. Since the latter wavelength corresponds to a photon energy of $800meV$, the $Si_{1-x}Ge_x$/Si heterosystem is inadequate, since a maximum practical valence band offset of only of the order of 500 meV can be obtained for $x = 0.5 \sim 0.6$, for Si layers sufficiently thin not to exceed the critical thickness. Therefore, alternate large bandgap, nearly lattice matched, barrier materials must be sought with sufficiently large band offsets with respect to silicon. Possible candidates include ZnS, CaF_2, SiO_2 or the Si/SiO_2 superlattice, and γ-Al_2O_3, among others[5]-[7].

The Si/ZnS heterosystem has received the most attention as current advances in epitaxy technology have allowed the growth of heterostructures consisting of polar and nonpolar materials[8, 9]. The lattice mismatch of ZnS with respect to Si is only 0.3%. The valence band offset has been predicted theoretically and experimentally[10]-[13]. Values range between 700 and 1900 meV, sufficiently large to give intersubband energy differences in the desired range. Growth of ZnS upon Si and Si upon ZnS have been demonstrated[9], with the use of an As monolayer to satisfy the local bonding requirements although the affect of the monolayer on the offsets has not been determined.

The possibility of population inversion and the operation of the intersubband laser depend critically on the lifetimes of the involved subbands. The subband lifetimes in turn are determined by nonradiative phonon scattering processes. The purpose of the present paper is to study the optical phonon modes in the Si/ZnS system since their interaction with carriers is considered to be dominant in the phonon scattering processes. This combination of materials is new, since it consists of both a nonpolar and polar semiconductor. Previous studies in carrier scattering by confined optical phonons in heterostructures have been focused only on one type of phonons, either polar[14]-[22] and nonpolar[23]-[26]. In the current situation involving both polar and nonpolar materials, carrier scattering by both types of phonons needs to be considered. To the best of our knowledge, there has not been any reported investigation on this mixed nature of optical phonons, their confinement effect, and their interaction with carriers in a heterostructure. In this paper, we will present a theoretical study based on the macroscopic continuum model to describe the confined optical phonon modes in a heterostructure consisting of polar and nonpolar materials. The ultimate intersubband laser design will likely consist of many periods, each of which will consist of more than one Si quantum wells coupled by ZnS barriers, with each period engineered to achieve

population inversion. However, in this initial investigation, we will consider only a simple superlattice consisting of alternating layers of Si and ZnS. The results of this study will provide the basis for the more complex structure described above.

As described below in greater detail, since the optical dispersions (frequency versus wavevector) of the silicon (Si) and zinc sulphide (ZnS) have no overlap, the optical phonons are assumed to be totally confined in both materials. In the silicon layers, a continuum model with double hybridization of the longitudinal optical (LO) and transverse optical (TO) modes is used to describe the vibration patterns of the guided modes[23]. The only boundary condition that needs to be satisfied in the Si layers is the vanishing of the displacements at the Si-ZnS interface, since the ZnS layers can be considered as infinitely rigid with respect to the vibrations of the Si layer. Hence, there is no interface mode in the Si layers. The situation on the ZnS layers is more complex. Following the work by Ridley[15, 16], here a continuum model is employed with hybridization of the optical LO, TO, and interface polariton (IP) modes needed to satisfy both the mechanical and electrostatic boundary conditions at the interfaces. Specifically, the electrostatic boundary conditions are the continuity of E_x, the electric field parallel to the interface, and the continuity of D_z, the displacement field normal to the interface. The mechanical boundary condition is again the vanishing of the optical displacements since the Si layers can be considered as infinitely rigid with respect to the vibrations of the ZnS layers.

Our current work provides a complete set of analytical expressions for the optical phonon dispersion relations, optical displacements, and associated scalar and vector potentials. These expressions can be used directly in calculating the interaction of carriers with the confined optical phonons.

II. Mode Patterns and Dispersion Relationship

A continuum model for the optical modes in the Si/ZnS superlattice is employed. Both mechanical and electrical boundary conditions are satisfied at the heterointerfaces. Since the optical dispersion relations (frequency versus phonon wavevector) in the two bulk materials have no overlap, the phonons are taken to be confined in their respective materials. For the Si layers, the continuum model for optical phonons in nonpolar materials[23, 25] is used. Here double hybridization of the LO (longitudinal optical) and TO (transverse optical) modes is used to give the vibration patterns of the guided modes. Since the ZnS layers are infinitely rigid with respect to the vibrations of the Si layers, only the mechanical boundary condition, the vanishing of the displacements at the interfaces, has to be satisfied.

For the polar ZnS layers, an alternate continuum model developed by Ridley and coworkers[15, 16] is employed. The situation is more complex than for nonpolar materials. Here, in order to satisfy both the electrostatic and mechanical boundary conditions, an intermixing of confined LO, TO, and IP (interface polariton) modes is needed. The boundary conditions which must be satisfied are (1) the continuity of E_x,

the component of electric field parallel to the interface, (2) the continuity of D_z, the component of the displacement vector normal to the interface, and (3) the vanishing of the vector displacement u at the interface.

A. Modes in Si Layers

As discussed above, since the ZnS layers can be treated as infinitely rigid, the boundary condition to be satisfied in the Si layers is the vanishing of the ionic displacement of all confined vibration modes. This is an assumption of strict confinement yielding only the guided modes. As pointed out in the continuum theory[23], the ionic displacement of confined vibrations has two components: one is the hybrid of the LO and p-polarized TO (p-TO) modes, and other is the uncoupled s-polarized TO (s-TO) mode. These modes are defined as follows: If we consider a (x, z) plane containing the normal to the layers and the phonon wavevector $\mathbf{Q}$, then

$$\mathbf{Q} = q_x \hat{e_x} + q_z \hat{e_z} \tag{1}$$

where $\hat{e_x}$ and $\hat{e_z}$ are unit vectors. The p-TO mode has its displacements normal to $\mathbf{Q}$ and in the plane, while the s-TO displacements are normal to $\mathbf{Q}$ and perpendicular to the plane ($||\hat{e_y}$).

The form of the ionic displacement, scalar, and vector potentials in one superlattice period differs from that in a neighboring period only by a phase factor proportional to the Bloch superlattice wavevector q_{SL}. Their expressions given below are obtained by taking $q_{SL} = 0$. A description of the s-TO mode is

$$u_y = e^{iq_x x}(A_{s-TO}e^{iq_z z} + B_{s-TO}e^{-iq_z z}), \tag{2}$$

while the hybrid of the LO and p-TO modes is given by

$$\begin{aligned} u_x &= \epsilon^{iq_x x}[q_x(A_{LO}e^{iq_L z} + B_{LO}e^{-iq_L z}) + q_T(A_{p-TO}e^{iq_T z} + B_{p-TO}e^{-iq_T z})], \\ u_z &= \epsilon^{iq_x x}[q_L(A_{LO}e^{iq_L z} - B_{LO}e^{-iq_L z}) - q_x(A_{p-TO}e^{iq_T z} - B_{p-TO}e^{-iq_T z})]. \end{aligned} \tag{3}$$

which are confined within the Si layer with a width of d_{Si}, $0 < z < d_{Si}$. The z-components of the LO and TO wavevector have been distinguished by q_L and q_T, respectively.

Since the LO and TO modes must have the same frequency to be effectively coupled, we must satisfy the condition

$$\omega^2 = \omega_0^2 - \beta_L^2(q_x^2 + q_L^2) = \omega_0^2 - \beta_T^2(q_x^2 + q_T^2), \tag{4}$$

where β_L and β_T are the velocities of LO and TO dispersions in Si, respectively.

Using the boundary condition that $\mathbf{u} = 0$ at the interfaces gives for the s-TO mode

$$u_y = Ae^{iq_x x}\sin(q_z z), \qquad \text{with } q_z = \frac{n\pi}{d_{Si}} \tag{5}$$

where $n = 1, 2, \cdots$ and A is a mode coefficient. This mode does not mix with other modes.

The hybrid LO and p-TO modes admit two classes of solutions. The 'sine' solution is

$$\begin{aligned} u_x &= 2Be^{iq_x x} q_x [\cos(q_L z) - \cos(q_T z)], \\ u_z &= 2iBe^{ik_z x} [q_L \sin(q_L z) + \frac{q_x^2}{q_T} \sin(q_T z)], \end{aligned} \tag{6}$$

and the 'cosine' solution is

$$\begin{aligned} u_x &= 2iBe^{iq_x x} [q_x \sin(q_L z) + \frac{q_L q_T}{q_x} \sin(q_T z)], \\ u_z &= 2Be^{iq_x x} q_L [\cos(q_L z) - \cos(q_T z)] \end{aligned} \tag{7}$$

where

$$q_L = \frac{n_L \pi}{d_{Si}} \quad \text{and} \quad q_T = \frac{n_T \pi}{d_{Si}}, \tag{8}$$

where $n_L = 1, 2, \cdots$, $n_T = 3, 4, \cdots$, $n_T - n_L = 2, 4, 6, \cdots$, and B is a mode coefficient. No interface modes exist in the Si layer because of the boundary condition $\mathbf{u} = 0$.

B. Modes in ZnS Layers

The boundary conditions are the continuity of E_x, D_z, and the vanishing of $\mathbf{u}$ at the interfaces. These conditions can be satisfied by a unique linear combination of LO, TO, and IP modes with common frequency and common in-plane wavevector q_x.

$$\mathbf{u} = \mathbf{u}_{LO} + \mathbf{u}_{TO} + \mathbf{u}_{IP} \tag{9}$$

We will use this hybrid expression to calculate the electrical interaction with carriers which is considerably stronger than the optical deformation potential interaction. We need consider only the displacements u_x and u_z, since u_y associated with the s-TO mode has no related electric field and therefore does not interact with carriers electrically. Once again, the expressions are obtained by taking the Bloch superlattice wavevector $q_{SL} = 0$.

For the LO mode, the ionic displacement

$$\begin{aligned} u_x &= e^{i(q_x x - \omega t)} q_x (A_L e^{iq_L z} + B_L e^{-iq_L z}), \\ u_z &= e^{i(q_x x - \omega t)} q_L (A_L e^{iq_L z} - B_L e^{-iq_L z}) \end{aligned} \tag{10}$$

which is confined within the ZnS layer with a width of d_{ZnS}, $-d_{ZnS}/2 < z < d_{ZnS}/2$

The associated electric fields are

$$E_x = -\rho_o u_x, \qquad E_z = -\rho_o u_z, \tag{11}$$

where

$$\rho_o = \frac{e^*}{\epsilon_o \Omega}, \tag{12}$$

with the effective ionic charge

$$e^{*2} = M\Omega\omega_{LO}^2\epsilon_o^2(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_s}) \tag{13}$$

where M is the reduced mass, ϵ_o is the permittivity of free space, ϵ_∞, ϵ_s are the high-frequency and static permittivities, and Ω is the volume of primitive unit cell. The scalar potential ϕ associated with the electric field $\mathbf{E} = -\nabla\phi$ is in turn given as

$$\phi = -i\rho_o e^{i(q_x x - \omega t)}(A_L e^{iq_L z} + B_L e^{-iq_L z}), \tag{14}$$

For the TO mode

$$\begin{aligned} u_x &= e^{i(q_x x - \omega t)} q_T (A_T e^{iq_T z} + B_T e^{-iq_T z}), \\ u_z &= -e^{i(q_x x - \omega t)} q_L (A_T e^{iq_T z} - B_T e^{-iq_T z}) \end{aligned} \tag{15}$$

The electric fields associated with this mode are negligible.

For the IP mode

$$\begin{aligned} u_x &= e^{i(q_x x - \omega t)} q_p (A_P e^{iq_p z} + B_P e^{-iq_p z}), \\ u_z &= ie^{i(q_x x - \omega t)} q_p (A_P e^{iq_p z} - B_P e^{-iq_p z}) \end{aligned} \tag{16}$$

The associated electric fields are

$$E_x = -\rho_p u_x, \qquad E_z = -\rho_p u_z, \tag{17}$$

where

$$\rho_p = \rho_o \frac{\omega^2 - \omega_{TO}^2}{\omega_{LO}^2 - \omega_{TO}^2}. \tag{18}$$

The electric fields associated with the interface modes propogate into the Si layers although they are treated as infinitely rigid and do not contain ZnS ionic displacement.

Being a transverse electromagnetic wave, there is a vector potential $\mathbf{A}$ associated with the electric field $\mathbf{E} = -\partial\mathbf{A}/\partial t$. Within the ZnS layers,

$$\begin{aligned} A_x &= i\frac{\rho_p}{\omega} e^{i(q_x x - \omega t)} q_p (A_P e^{iq_p z} + B_P e^{-iq_p z}), \\ A_z &= -\frac{\rho_p}{\omega} e^{i(q_x x - \omega t)} q_p (A_P e^{iq_p z} - B_P e^{-iq_p z}) \end{aligned} \tag{19}$$

While in the Si layers, a similar expression can be obtained with another set of mode coefficients, A_{p1} and B_{p1}.

Under the assumption of long wavelength waves and elastically isotropic medium, the requirement for common frequency gives the dispersion relationship,

$$\begin{aligned} \omega^2 &= \omega_{LO}^2 - v_L^2(q_x^2 + q_L^2) \\ &= \omega_{TO}^2 - v_T^2(q_x^2 + q_T^2) \\ &= \frac{c^2(q_x^2 + q_p^2)}{\epsilon(\omega)\mu_o} \end{aligned} \tag{20}$$

where v_L and v_T are velocities approximately equal to those of LA and TA modes in ZnS, respectively, c is the velocity of light in vacuum and μ_o is the permittivity of free space. In the above expressions, the frequency in the ZnS layers lies between the ZnS LO and TO zone center frequencies. Since $\omega_{TO} < \omega_{LO}$, in order for the TO frequency to be equal to a LO frequency q_T must be imaginary $q_T = iq_o$, corresponding to a TO interface mode. The modes which interact most strongly with carriers are those with frequencies near the LO branch. For these modes, the value of q_o is large, and we can take the approximation

$$\tanh(q_o d_{ZnS}) \approx 1. \tag{21}$$

In the unretarded limit ($c \to \infty$), $q_x^2 + q_p^2 \approx 0$ for the IP mode. Hence, $q_p \approx iq_x$.

Applying, at the two interfaces between layers Si and ZnS in a period of the superlattice, the conditions that u_x and u_z equal to zero along with the continuity of E_x and D_z, leads to eight simultaneous equations involving the eight unknown mode coefficients ($A_L, B_L; A_T, B_T; A_P, B_P$; and A_{P1}, B_{P1}). The following two ionic displacement mode patterns emerge for the hybrid in Eq.(9) taking the Bloch superlattice wavevector $q_{SL} = 0$ and the approximation $\tanh(q_o d_{ZnS}) \approx 1$. For the first type,

$$\begin{aligned}
u_x = \; & 2iBe^{iq_x x} q_x \{\sin(q_L z) \\
& -[1 - p_1 \tanh(q_x d_{ZnS}/2)] \sin(q_L d_{ZnS}/2) \frac{\sinh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\
& -p_1 \sin(q_L d_{ZnS}/2) \frac{\sinh(q_x z)}{\cosh(q_x d_{ZnS}/2)}\}, \\
u_z = \; & 2Be^{iq_x x} q_L \{\cos(q_L z) \\
& -\frac{q_x^2}{q_L q_o}[1 - p_1 \tanh(q_x d_{ZnS}/2)] \sin(q_L d_{ZnS}/2) \frac{\cosh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\
& -\frac{q_x}{q_L} p_1 \sin(q_L d_{ZnS}/2) \frac{\cosh(q_x z)}{\cosh(q_x d_{ZnS}/2)}\},
\end{aligned} \tag{22}$$

and for the second type,

$$\begin{aligned}
u_x = \; & 2Be^{iq_x x} q_x \{\cos(q_L z) \\
& -[1 - p_2 \coth(q_x d_{ZnS}/2)] \cos(q_L d_{ZnS}/2) \frac{\cosh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\
& -p_2 \cos(q_L d_{ZnS}/2) \frac{\cosh(q_x z)}{\sinh(q_x d_{ZnS}/2)}\}, \\
u_z = \; & 2iBe^{iq_x x} q_L \{\sin(q_L z) \\
& +\frac{q_x^2}{q_L q_o}[1 - p_2 \coth(q_x d_{ZnS}/2)] \cos(q_L d_{ZnS}/2) \frac{\sinh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\
& +\frac{q_x}{q_L} p_2 \cos(q_L d_{ZnS}/2) \frac{\sinh(q_x z)}{\sinh(q_x d_{ZnS}/2)}\},
\end{aligned} \tag{23}$$

where

$$
\begin{aligned}
p_1 &= -\frac{\alpha+\gamma}{2rsd}, \\
p_2 &= \frac{\gamma-\alpha}{2rsd}, \\
\alpha &= \sinh(q_x d_{Si})\cosh(q_x d_{ZnS}) + r\cosh(q_x d_{Si})\sinh(q_x d_{ZnS}), \\
\gamma &= \sinh(q_x d_{Si}) + r\sinh(q_x d_{ZnS}), \\
d &= 1 - Z\sinh(q_x d_{ZnS})\sinh(q_x d_{Si}) - \cosh(q_x d_{ZnS})\cosh(q_x d_{Si}), \\
Z &= \frac{1}{2}(r+\frac{1}{r}), \\
s &= \frac{\omega^2-\omega_{TO}^2}{\omega_{LO}^2-\omega_{TO}^2}, \\
r &= \frac{\epsilon_{p1}}{\epsilon_{p2}}
\end{aligned}
\tag{24}
$$

where ϵ_{p1} and ϵ_{p2} are the permittivities in Si and ZnS layers, respectively, with

$$
\epsilon_{p2} = \epsilon_\infty \frac{\omega^2-\omega_{LO}^2}{\omega^2-\omega_{TO}^2}. \tag{25}
$$

C. Dispersion Relationship

The phonon frequency in the ZnS layers is determined by the following set of equations:

$$
\begin{cases}
\omega^2 = \omega_o^2 - v_L^2(q_x^2+q_L^2), \\
\omega^2 = \omega_o^2 - v_T^2(q_x^2-q_o^2), \\
t_1 + t_2\cos(q_L d_{ZnS}) + t_3\sin(q_L d_{ZnS}) = 0
\end{cases}
\tag{26}
$$

where

$$
\begin{aligned}
t_1 &= 4p\sinh(q_x d_{Si}) + 4pr\sinh(q_x d_{ZnS}), \\
t_2 &= -4p\alpha, \\
t_3 &= 8p^2 r\sinh(q_x d_{ZnS})\sinh(q_x d_{Si}) - 4p^2\alpha^2 \\
&\quad +4p^2 r^2\sinh^2(q_x d_{ZnS}) + 4p^2\sinh^2(q_x d_{Si}) + 1,
\end{aligned}
\tag{27}
$$

and

$$
p = \frac{q_x}{4q_L rsd}. \tag{28}
$$

The third equation in (26) is obtained from the requirement of a nonzero solution for the eight simultaneous equations discussed above, and Eq.(27) is arrived under the approximation, $\tanh(q_o d_{ZnS}) \approx 1$.

The numerical procedure for determining a phonon frequency is the following: given a value of q_x, we can determine those of t_1, t_2, and t_3 from Eq.(27). Then ω is scanned from ω_{TO} to ω_{LO}. For a given value of ω, q_L and q_o are obtained from the first two equations in (26). Those values are then substituted into the third equation in (26) to determine if the particular value of ω is a solution.

III. Scalar and Vector Potentials

The study of optical modes in the Si/ZnS superlattice will be ultimately applied to calculate the electrical interaction between the optical phonons and carriers in the superlattice. For this purpose, expressions for the scalar and vector potentials are essential in obtaining the electrical interaction Hamiltonian with respect to such an interaction,

$$H = -e\phi + \frac{e}{m}\mathbf{A}\cdot\mathbf{p}, \tag{29}$$

where $\mathbf{p}$ is the momentum operator, e and m are the free electron charge and mass, respectively. The scalar potential ϕ associated with the LO mode vanishes in Si layers where there is only the $\mathbf{A}\cdot\mathbf{p}$ interaction.

Associated with the two types of ionic displacement in Eqs.(22) and (23), the scalar potentials in ZnS layers are given as, for the first type,

$$\phi = \begin{cases} 2\rho_o B e^{iq_x x}\sin(q_L z_1) & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ 0 & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer.} \end{cases} \tag{30}$$

and for the second type,

$$\phi = \begin{cases} -2i\rho_o B e^{iq_x x}\cos(q_L z_1) & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ 0 & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer.} \end{cases} \tag{31}$$

Note that we have used two different coordinates z_1 and z_2 for layers ZnS and Si, respectively, with their origins placed at the centers of the respective layers.

The vector potentials can be obtained, for the first type,

$$A_x = \begin{cases} \dfrac{2s\rho_o q_x}{\omega} B e^{iq_x x} p_1 \sin(q_L d_{ZnS}/2)\dfrac{\sinh(q_x z_1)}{\cosh(q_x d_{ZnS}/2)} & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ -\dfrac{4q_x\rho_o}{\omega d} B e^{iq_x x} V_1 \sinh(q_x z_2) & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \tag{32}$$

$$A_z = \begin{cases} \dfrac{2is\rho_o q_x}{\omega} B e^{iq_x x} p_1 \sin(q_L d_{ZnS}/2)\dfrac{\cosh(q_x z_1)}{\cosh(q_x d_{ZnS}/2)} & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ -\dfrac{4iq_x\rho_o}{\omega d} B e^{iq_x x} V_1 \cosh(q_x z_2) & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \tag{33}$$

and for the second type,

$$A_x = \begin{cases} -\dfrac{2is\rho_o q_x}{\omega} B e^{iq_x x} p_2 \cos(q_L d_{ZnS}/2)\dfrac{\cosh(q_x z_1)}{\sinh(q_x d_{ZnS}/2)} & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ -\dfrac{4iq_x\rho_o}{\omega d} B e^{iq_x x} V_2 \cosh(q_x z_2) & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \tag{34}$$

$$A_z = \begin{cases} -\dfrac{2s\rho_o q_x}{\omega} B e^{iq_x x} p_2 \cos(q_L d_{ZnS}/2) \dfrac{\sinh(q_x z_1)}{\sinh(q_x d_{ZnS}/2)} & |z_1| < \dfrac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ -\dfrac{4q_x \rho_o}{\omega d} B e^{iq_x x} V_2 \sinh(q_x z_2) & |z_2| < \dfrac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \tag{35}$$

where

$$\begin{aligned} V_1 = \; & \sin(q_L d_{ZnS}/2)\cosh(q_x d_{ZnS}/2) \\ & [\cosh(q_x d_{ZnS}/2)\sinh(q_x d_{Si}/2) + r\sinh(q_x d_{ZnS}/2)\cosh(q_x d_{Si}/2)], \\ V_2 = \; & \cos(q_L d_{ZnS}/2)\sinh(q_x d_{ZnS}/2) \\ & [\sinh(q_x d_{ZnS}/2)\cosh(q_x d_{Si}/2) + r\cosh(q_x d_{ZnS}/2)\sinh(q_x d_{Si}/2)]. \end{aligned} \tag{36}$$

IV. Results and Discussion

A. Mode Patterns in Si Layers

The lowest s-TO mode pattern in Eq.(5) for $q_z = \pi/d_{Si}$ is shown in Fig.1(a) within a Si layer of $d_{Si} = 40\text{Å}$, while the hybrid patterns of the lowest p-TO and LO modes with $q_L = \pi/d_{Si}$ and $q_T = 3\pi/d_{Si}$ are shown in Figs.1(b) and 3(c) for the 'sine' and 'cosine' solutions given in Eqs.(6) and (7), respectively within the same Si layer. The strict confinement which requires the vanishing of ionic displacements at the boundaries of Si layers is clearly demonstrated for both vibration modes.

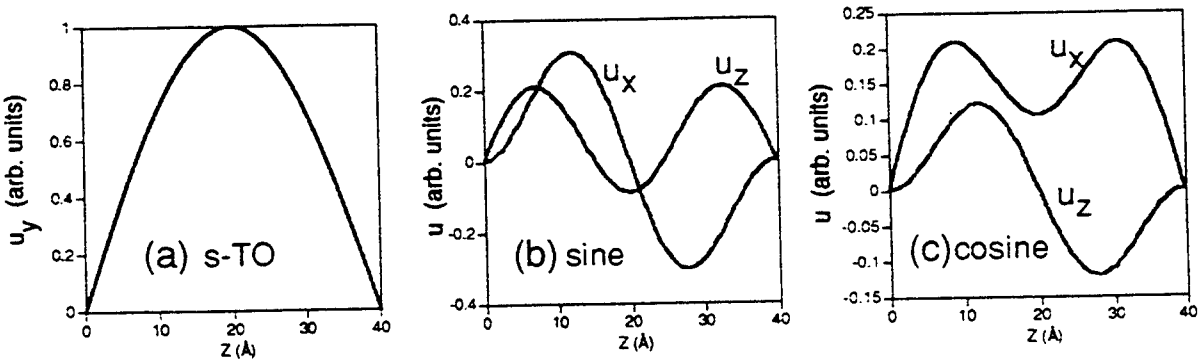


Figure 1. Vibration patterns in a Si layer with a width of 40Å for (a) the guided s-TO mode, (b) the 'sine' solution, and (c) the 'cosine' solution of the guided p-TO and LO modes

B. Mode Patterns in ZnS layers

To illustrate the patterns of ionic displacements in the ZnS layers given in Eqs.(22) and (23), we need to first determine values for q_x, q_L, and q_o. To do so, we will follow the numerical procedure described in Section II(C) by arbitrarily fixing a value for the in-plane phonon wavevector $q_x = \pi/(10a_{ZnS})$, where a_{ZnS} being the lattice constant of ZnS. This choice of q_x satisfies the requirement for in-plane wavevectors to be considered

large enough to neglect the effect of retardation so that $q_p = iq_z$, and also at the same time small enough so that the quadratic dispersion assumed for the LO and TO modes in Eq.(20) is valid. In the event of calculating the carrier-optical phonon interaction, the value of q_z is actually determined by the conservation of in-plane momentum between the initial and final states of the scattering process. For a given value of q_z, typically, a set of hybridized modes can be obtained. Here, we show only the mode pattern with frequency close to ω_{LO}.

We obtained $\hbar\omega = 35.5meV$, $q_L = 0.31 \times 10^8/cm$ and $q_o = 0.98 \times 10^8/cm$. Fitting these values into Eqs.(22) and (23), we obtained Figs.2(a) and 2(b) showing the mode patterns of ionic displacement of both the first and second types, respectively in a ZnS layer of $d_{ZnS} = 20\mathring{A}$. It can be seen from Figs.2(a) and 2(b) that the mechanical boudary condition, vanishing of the ionic displacements at the interfaces of Si and ZnS layers, is satisfied.

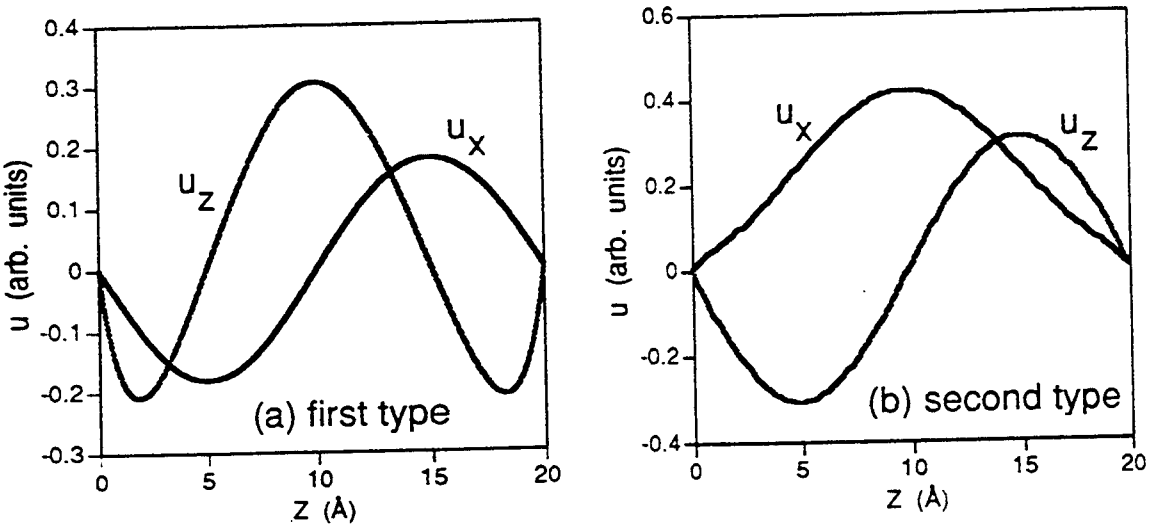


Figure 2. Vibration patterns in a ZnS layer with a width of 20Å for (a) the first type and (b) the second type solutions of the hybridized LO, TO and IP modes.

C. Potential and Field Distributions in the Superlattice

The scalar potentials associated with the LO modes are strictly confined within the ZnS layers. Their distributions are shown in Fig.3 for the first and second types given in Eqs.(30) and (31) with $q_L = 0.31 \times 10^8/cm$, $d_{ZnS} = 20\mathring{A}$, respectively.

The vector potential associated with the IP modes are distributed in both Si and ZnS layers, even thought Si layers are treated as infinitely rigid and do not contain ZnS ionic displacements. The profiles for the two components of the vector potentials given in Eqs.(32-35) for the first and second types with $d_{Si} = 40\mathring{A}$, $d_{ZnS} = 20\mathring{A}$ are shown in Figs.4(a) and 4(b), respectively.

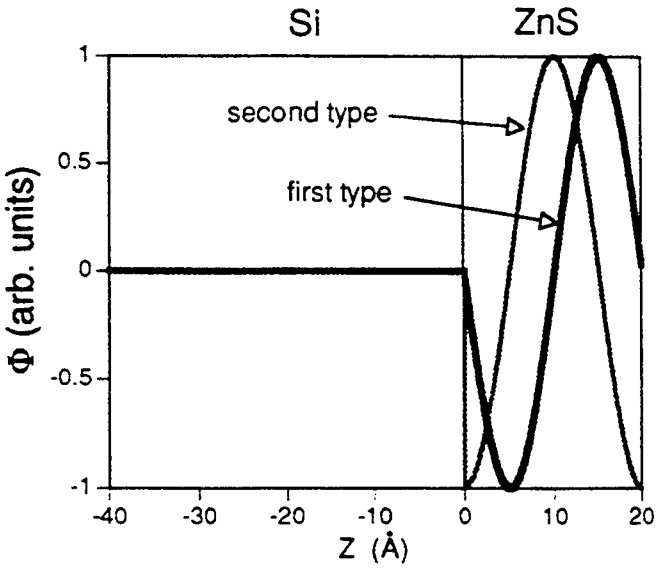


Figure 3. Scalar potential distribution associated with the LO modes in a period of the Si/ZnS superlattice with $d_{Si} = 40\mathring{A}$ and $d_{ZnS} = 20\mathring{A}$ for both the first and second types of the vibration modes.

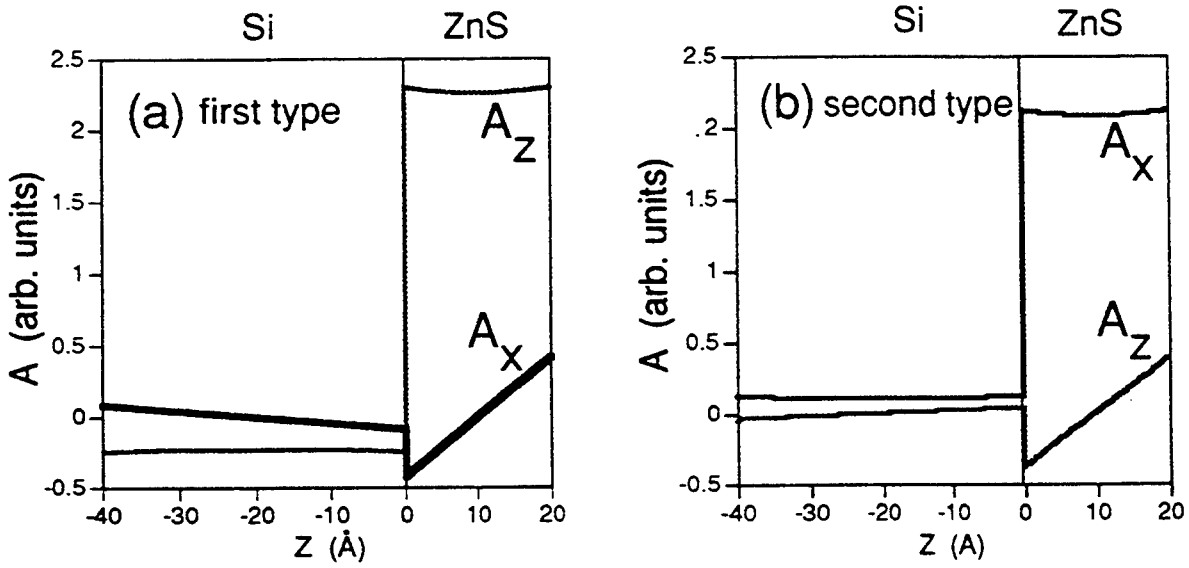


Figure 4. Vector potentials associated with the IP modes distributed in a period of the Si/ZnS superlattice with $d_{Si} = 40\mathring{A}$ and $d_{ZnS} = 20\mathring{A}$ for (a) the first type and (b) the second type of the vibration modes

It can be seen from Figs.3 and 4 that both scalar and vector potentials are not continuous across the interfaces. However, as pointed by Ridley[15], the energy of interaction with an electron traveling coherently with the optical phonon is continuous. The electric field can be obtained as

$$\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}. \tag{37}$$

The continuity of E_x and $D_z = \epsilon(\omega)E_z$ implies that at boundaries,

$$\begin{aligned} \omega A_x|_{z_2=\pm d_{Si}/2} &= -q_x\phi|_{z_1=\mp d_{ZnS}/2} + \omega A_x|_{z_1=\mp d_{ZnS}/2}, \\ A_z|_{z_2=\pm d_{Si}/2} &= r A_z|_{z_1=\mp d_{ZnS}/2}, \end{aligned} \tag{38}$$

where A_{1x} and A_{1z} are x- and z-components of the vector potential in Si layers. The interaction in the Si layer is $e(A_{1x}v_x + A_{1z}v_z)$ and in the ZnS layer $e(-\phi + A_x v_x + A_z v_z)$, which are equal when the electron velocity $v_x = \omega/q_x$ and $v_z = 0$. Thus, the coherent interaction energy is continuous across the interfaces.

The electric field distributions for E_x and $\epsilon(\omega)$ in Si ($d_{Si} = 40\text{Å}$) and ZnS ($d_{ZnS} = 40\text{Å}$) layers are shown in Figs.5(a) and 5(b) for the first and second types, respectively. The continuity of E_x and D_z across the Si and ZnS interface according to Eq.(38) is clearly demonstrated.

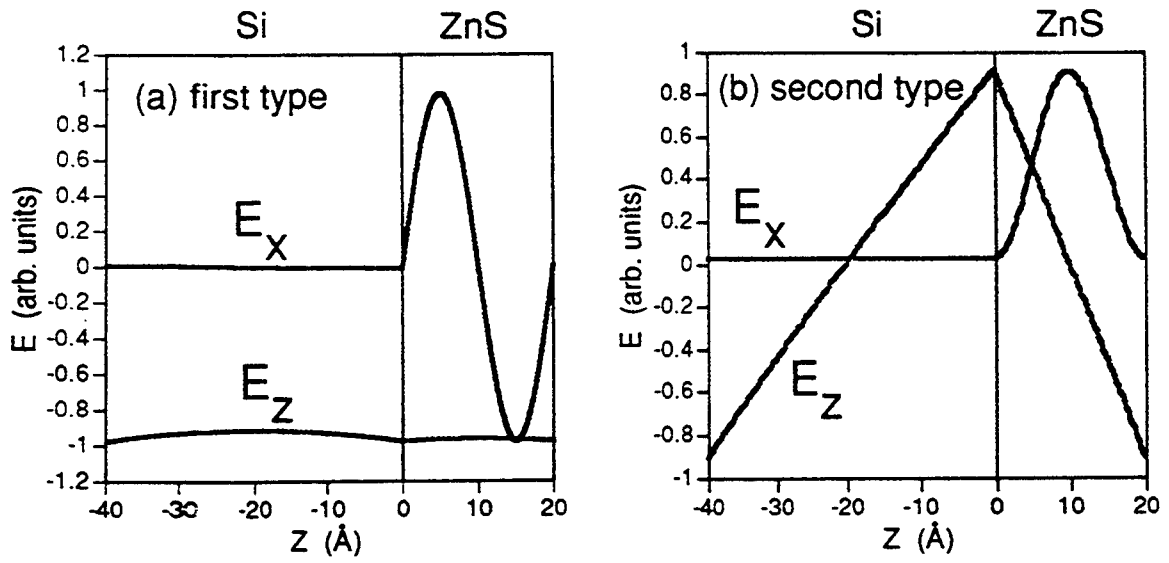


Figure 5. The field distributions, E_x and D_z, derived from the scalar and vector potentials. in a period of the Si/ZnS superlattice with $d_{Si} = 40\text{Å}$ and $d_{ZnS} = 20\text{Å}$ for (a) the first type and (b) the second type of the vibration modes.

V. Conclusions

We have provided an analytical model of optical modes in Si/ZnS superlattices consisting of polar and nonpolar optical phonons. In the Si layers, a continuum model with double hybridization of the LO and TO modes is used to described the vibration patterns. Since there is no electric field resulted from the nonpolar ionic displacements in Si layers, the only boundary condition that needs to be satisfied in the Si layers is the vanishing of the displacements at the Si-ZnS interface as the ZnS layers can be considered as infinitely rigid with respect to the vibrations of the Si layer. Due to this strict confinement, only guided modes emerge in the Si layers which consist of s-TO and coupled p-TO and LO modes, with no interface modes. These guided modes have been illustrated. Their interaction with carriers in the superlattice can be calculated through the optical deformation potential for Si. The interaction Hamiltonian can actually be obtained by taking the product of this potential with the normalized ionic displacement.

However, for the optical phonons in ZnS layers, we need to include the electrical interaction in calculating the carrier scattering by optical phonons, since there are electric fields associated with the polar optical vibrations. As a result, both mechanical and electrostatic boundary conditions needs to be satisfied in the interfaces. A continuum model employing a linear combination of LO, TO and IP (interface polariton) modes with a common frequency is used to describe the ionic displacements in ZnS layers. A numerical procedure for determining a phonon frequency is provided. This hybridized model is necessary to meet the simultaneous requirement on the mechanical and electrostatic boundary conditions. The mechanical boundary condition is again the vanishing of the optical displacements since Si layers can be considered as infinitely rigid with respect to the vibrations of the ZnS layers. The electrostatic boundary conditions are the continuity of the electric field parallel to the interface, and the continuity of the displacement field normal to the interface. Based on this set of boundary conditions, expressions are obtained for the ionic displacements in ZnS layers consisting of LO, TO, and IP modes. There are scalar and vector potentials associated with the LO and IP modes, respectively, but no electric field associated with the TO mode. The scalar potential and its associated electric field due to the LO mode are distributed only within the ZnS layers and are zero in the Si layers. But the vector potential and its associated electric field due to the IP mode have distributions in both ZnS and Si layers even though there is no ZnS ionic displacement mode in the Si layers. Examples of these mode characteristics have been demonstrated. Neither the scalar nor vector potential is continuous across the Si-ZnS interface, but the energy of coherent interaction with carriers is continuous due to the continuity of the electric field parallel to the interface. The analytical model for the confined optical modes consisting of polar and nonpolar optical phonons will be employed in calculating the carrier-phonon interaction to estimate the subband lifetimes in the Si/ZnS superlattices.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Richard A. Soref for his support during the period of this summer research program, without which this research accomplishment would not have been possible. The author would also like to acknowledge Dr. Lionel Friedman for many helpful discussions.

References

[1] J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, C. Sirtory, and A. Y. Cho, Science **264**, 553 (1994)

[2] J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, C. Sirtory, S. N. G. Chu, and A. Y. Cho, Appl. Phys. Lett. **65**, 2091 (1994)

[3] G. Sun, L. Friedman, and R.A. Soref, Appl. Phys. Lett. **66**, 3425 (1995)

[4] R. A. Soref, Proc. IEEE **81**, 1687 (1993)

[5] L. Friedman and R. A. Soref, IEEE Photonics Technology Letters **5**, 1200 (1993)

[6] L. J. Schowalter and R. W. Fathauer, CRC Critical Review **15**, 367 (1989)

[7] R. Tsu, Nature **364**, 19 (1993)

[8] M Yokoyama, K. I. Kashiro, and S. I. Ohta, J. Crystal Growth **81**, 73 (1987)

[9] X. Zhou and W. P. Kirk, Mat. Res. Soc. Symp. Proc. **318**, 207 (1994)

[10] E. G. Wang and C. S. Ting, Phys. Rev. B **51**, 9791 (1995)

[11] C. Maierhofer, S. Kulkarni, M. Alonso, T. Reich, and K. Horn, J. Vac. Sci. Technol. B **9**, 2238 (1991)

[12] M. Cardona and N. E. Christensen, J. Vac. Sci. Technol. B **6**, 1285 (1988)

[13] W. A. Harrison, J. Vac. Sci. Technol. **14**, 1016 (1977)

[14] B. K. Ridley, Phys. Rev. B **39**, 5282 (1989)

[15] B. K. Ridley, Phys. Rev. B **47**, 4592 (1993)

[16] M. P. Chamberlain, M Cardona, and B. K. Ridley, Phys. Rev. B **48**, 14356 (1993)

[17] N. C. Constantinou and B. K. Ridley, Phys. Rev. B **49**, 17065 (1994)

[18] B. K. Ridley, Appl. Phys. Lett. **66**, 3633 (1995)

[19] E. Molinari and A. Fasolino, Appl. Phys. Lett. **54**, 1220 (1989)

[20] L. register. Phys. Rev. B **45**, 8756 (1992)

[21] K. J. Nash. Phys. Rev. B **46**, 7723 (1992)

[22] N. Mori and T. Ando, Phys. Rev. B **40**, 6175 (1989)

[23] G. Sun and L. Friedman, Phys. Rev. B **53**, 3966 (1995)

[24] A. Fasolino, E. Molinari, and J.C. Mann, Phys. Rev. B **39**, 3923 (1989)

[25] B. K. Ridley, Phys. Rev. B **44**, 9002 (1991)

[26] S. C. Jain and W. Hayes, Semicond. Sci. Technol. **6**, 547 (1991)